НИЯУ МИФИ

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ ЧАСТИЦ В СЦИНТИЛЛЯЦИОННЫХ ДЕТЕКТОРАХ

Выполнил студент М25-112
Козлов А.
Научный руководитель:
Попов Д.В.

Введение

Современная экспериментальная физика элементарных частиц активно расширяет свой арсенал инструментов для изучения окружающего мира: помимо ставших классическими методов анализа экспериментальных данных, разработанных в течение XX века, возникают новые решения на стыке физических и компьютерных наук.

Одним из таких методов является применение методов машинного обучения (Machine Learning или ML) в процессе сбора, обработки и анализа экспериментальных данных в различных экспериментах. Внедрение ML в физику началось относительно недавно, но уже принесло огромные результаты, сделав преимущества использования ИИ очевидными.

Таким образом, применение методов ML может позволить качественно улучшить результаты проводимых экспериментов и предложить новый подход к решению конкретных физических задач.

Цель работы

Изучение возможностей применения методов ML в рамках задачи классификации частиц в сцинтилляционных детекторах

Задачи:

- Изучить экспериментальную схему установки для получения данных
- Провести обработку данных, конструирование признаков (feature engineering)
- Исследовать работу нескольких классических алгоритмов ML в рамках эксперимента

Экспериментальная установка

Модель детектора типа Phoswich (Phosphorus sandwich)

- Быстрый пластиковый сцинтиллятор, соединённый с сцинтиллятором LCS (литий-кальций-силикат)
- Сцинтиллятор NaI для реализации схемы совпадений и задержанных совпадений
- Источник - калифорний-252: гамма-кванты и нейтроны

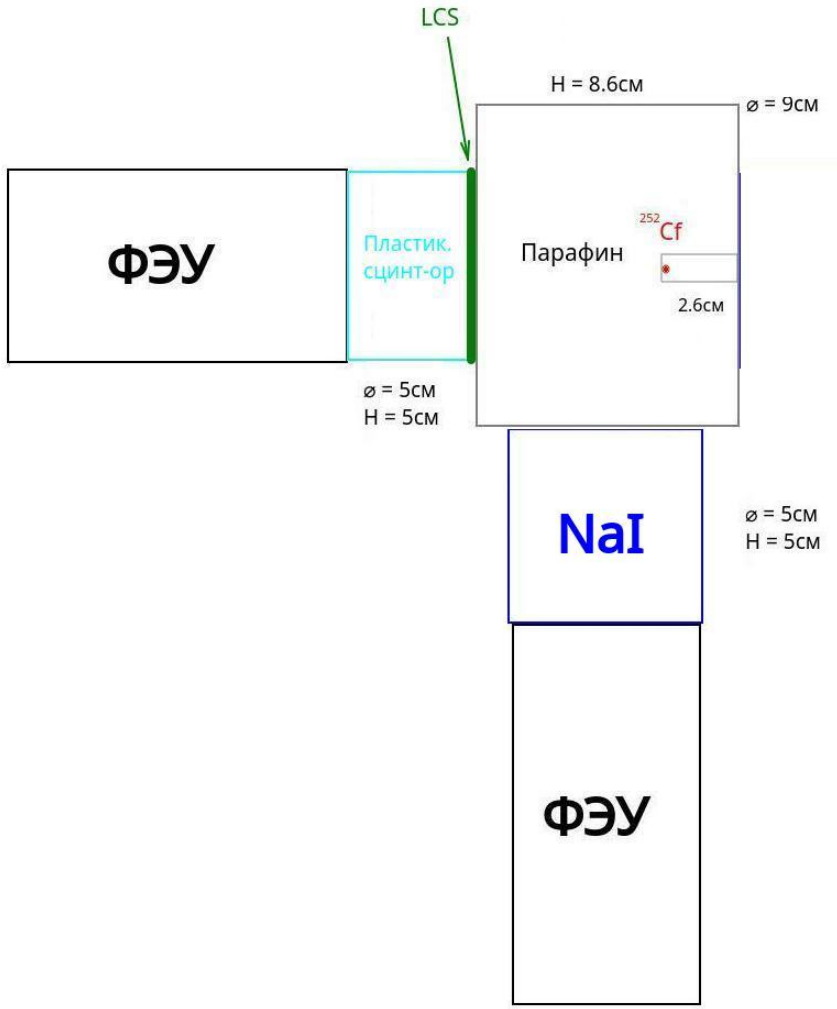


Рис. 1 Схема установки

Экспериментальные данные

Данные с детектора записывались поканально при помощи DAQ – таким образом, удавалось сохранить информацию о ключевых характеристиках сигнала

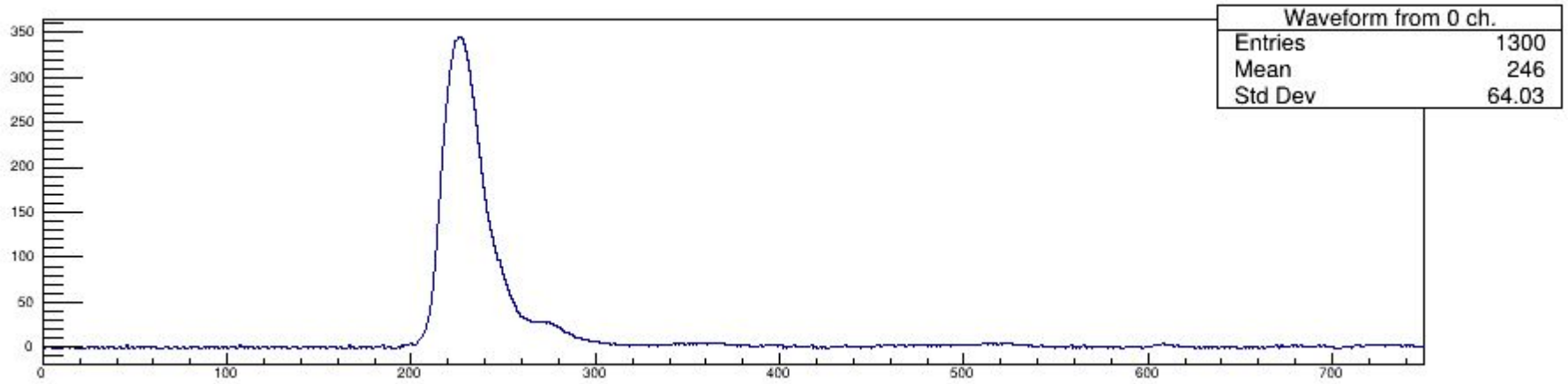


Рис. 2 Характерная форма записи информации о событии в DAQ

Далее при помощи ROOT-скрипта происходило извлечение дополнительных сведений о сигнале

Метод PSD

Метод PSD (Pulse Shape Discrimination) является классическим для задачи отделения нейтрон-гамма событий, однако он имеет ряд недостатков:

1. Произвол в выборе PSD-критерия
2. Сильная зависимость результата от электронной конфигурации эксперимента
3. Произвол в выборе промежутка интегрирования

Эти недостатки могут быть устранены при использовании методов ML

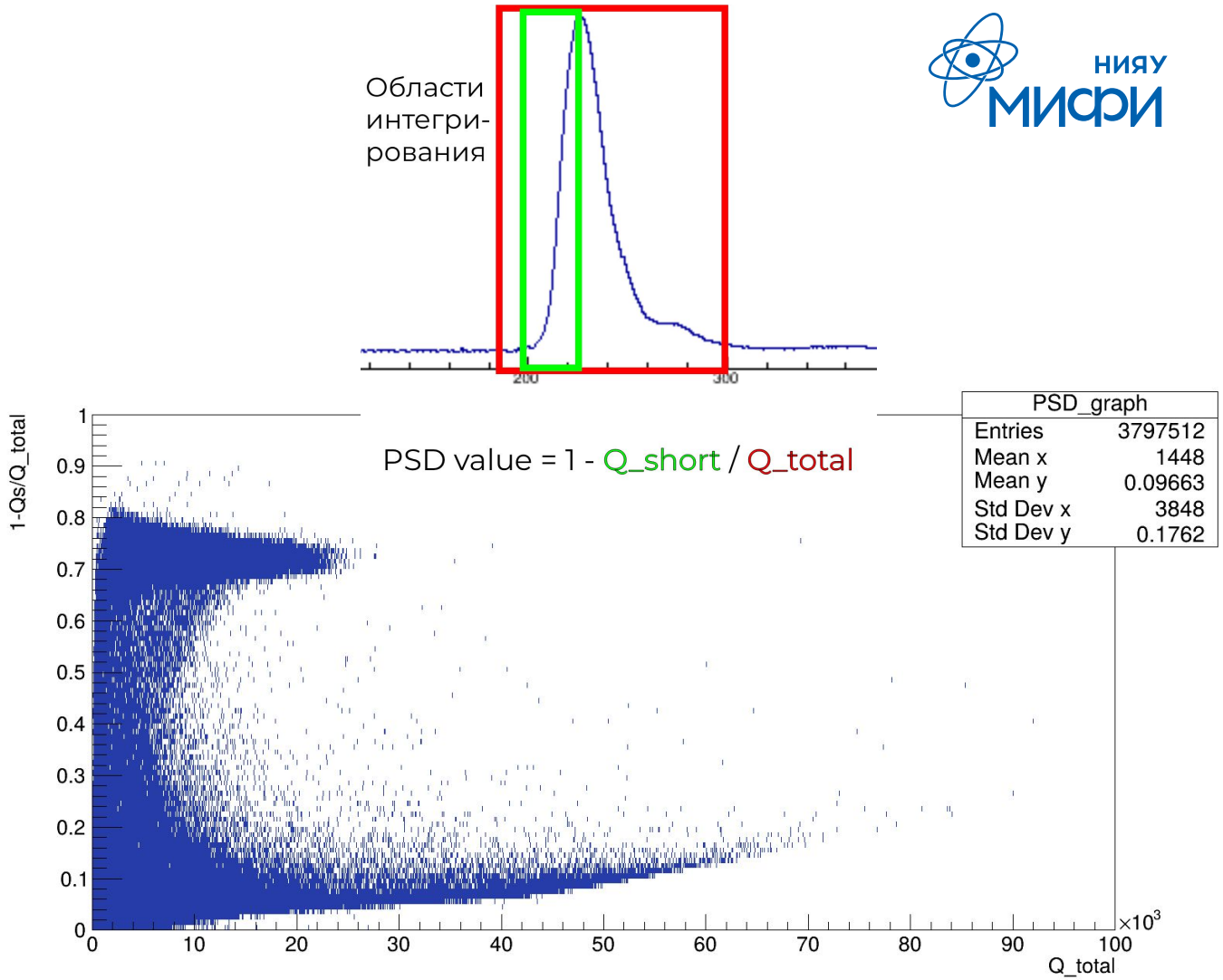


Рис. 3 Применение метода PSD для детектора Phoswich

Обработка данных и конструирование признаков

Данные эксперимента хранятся в ROOT-дереве после первичной обработки ROOT-скриптом

В качестве признаков для обучения моделей были использованы:

- Front Time – время нарастания переднего фронта
- Tail Time – время спада заднего фронта
- FWHM – полная ширина на полувысоте сигнала
- skewness (коэффициент асимметрии) – третий момент распределения
- kurtosis (эксцесс) – четвёртый момент распределения
- Q_total / Q_muon – нормированный полный заряд на заряд жёстких мюонов

В качестве признака намеренно не был взят PSD_value в силу своей нефизичности

Конструирование признаков для модели

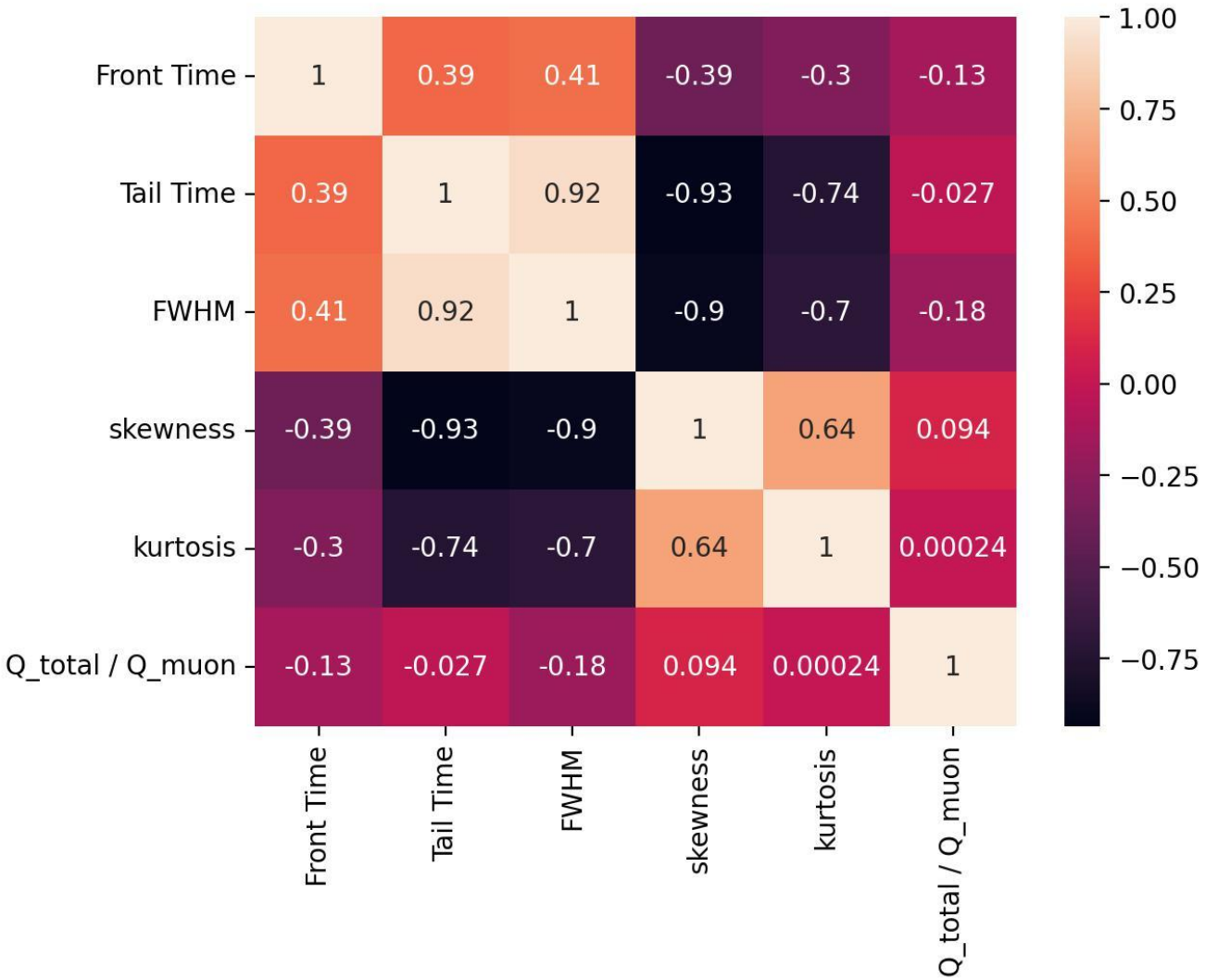


Рис. 4 Матрица корреляций выбранных признаков

Среди выбранных признаков наблюдаются пары с коэффициентом корреляции и антикорреляции близким к единице – это необходимо учитывать при построении модели

- Front Time – время нарастания переднего фронта
- Tail Time – время спада заднего фронта
- FWHM – полная ширина на полувысоте сигнала
- skewness (коэффициент асимметрии) – третий момент распределения
- kurtosis (эксцесс) – четвёртый момент распределения
- Q_total / Q_muon – нормированный полный заряд на заряд жёстких мюонов

Методика обучения моделей

1. В качестве обучающей выборки берутся данные эксперимента, соответствующие областям графика PSD с известным типом частиц: мюоны – жёсткий край спектра, гамма-кванты – мягкая часть спектра, малые PSD value, нейтроны – отбор по PSD value

2. Эти данные размечаются по типу частиц, затем разбиваются на train-test набор, который участвует в обучении модели и в оценке эффективности её обучения

3. При необходимости происходит масштабирование признаков и сохранение скейлера для дальнейшего использования модели

4. Производится обучение модели и оценка её эффективности

5. Также производится дополнительная валидация эффективности модели при помощи дополнительных экспериментальных данных с событиями гамма-квантов

Классификатор на основе модели логистической регрессии

Три класса для обучения модели:
1. Гамма-кванты
2. Нейтроны
3. Мюоны

Используемая библиотека: scikit-learn, метод LogisticRegression с L1-регуляризацией и классификатором One-vs-Rest (для 3 классов)

Достоинства метода:
- Простота реализации и расчётов
- Регуляризация параметров
- Результат: вероятность принадлежности к классу + интерпретируемые веса

$z=(w)^{\top}*X+b$

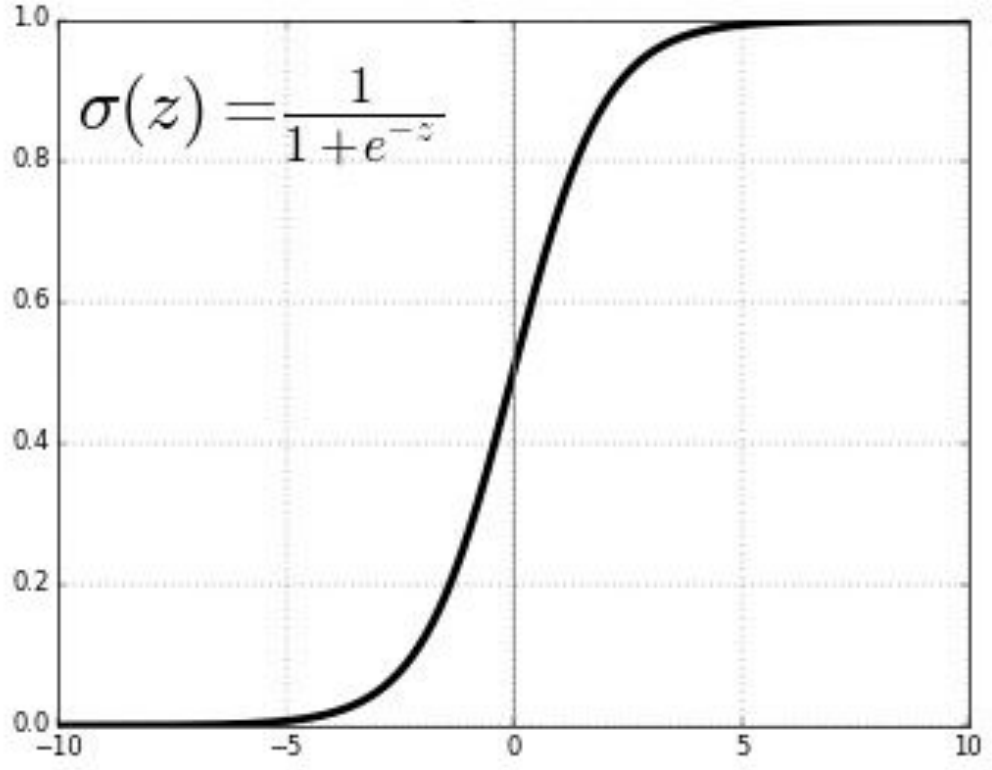


Рис. 5 Пример сигмоиды, используемой для классификации методом логистической регрессии

Классификатор на основе модели логистической регрессии

Результат использования логистической регрессии на тестовой выборке:

	precision	recall	f1-score	support
gamma	1.00	1.00	1.00	3830
muon	1.00	1.00	1.00	327
neutron	1.00	1.00	1.00	3439
accuracy			1.00	7596
macro avg	1.00	1.00	1.00	7596
weighted avg	1.00	1.00	1.00	7596

Результат валидации данных данными с источника гамма-квантов:

Accuracy = 0.93 при добавлении порога уверенности в 80% (при непреодолении порога частица классифицируется как “Not Classified”)

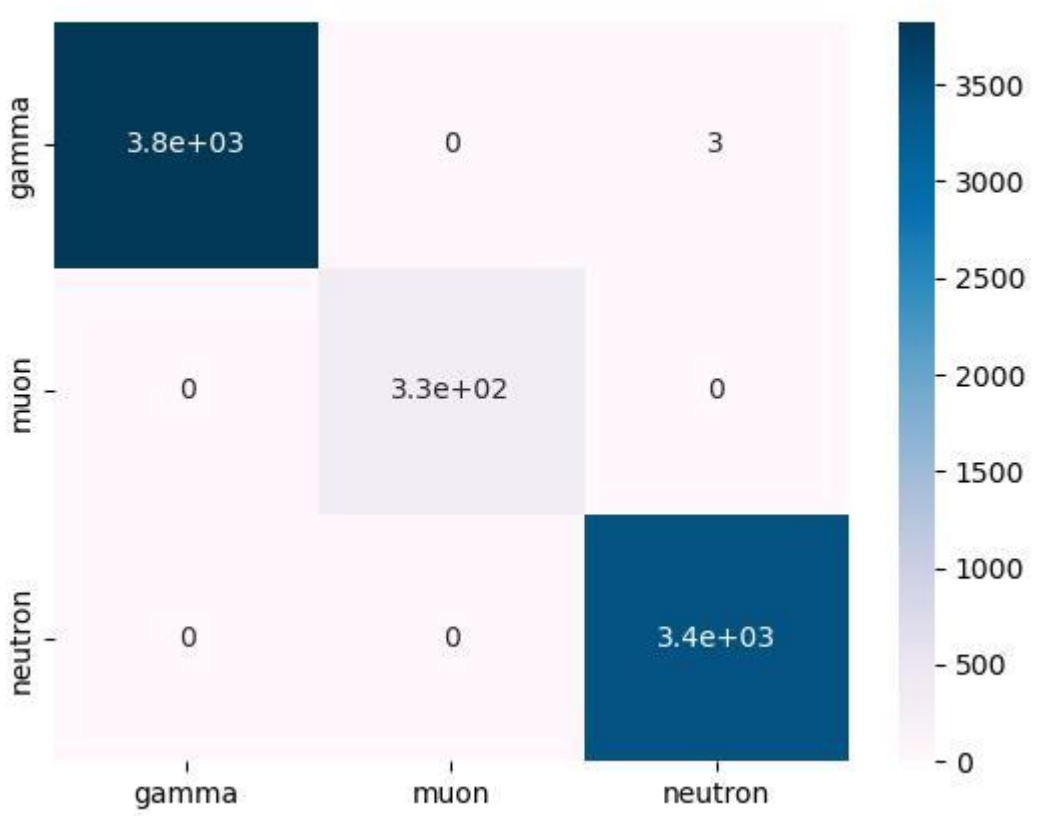


Рис. 6 Confusion-Matrix для модели логистической регрессии (по вертикали истинные значения, по горизонтали – предсказания модели)

Классификатор на основе модели логистической регрессии

Рис. 7 Разметка данных моделью логистической регрессии на графике PSD

Классификатор на основе BDT

BDT (Boosted Decision Tree) – другой классический алгоритм ML, часто применяемый в физических экспериментах в силу своих достоинств. Метод представляет собой ансамбль классических деревьев решения (Decision Tree)

Достоинства BDT:

- Хорошая интерпретируемость метода
- Возможность точно описать сложные нелинейные разбиения данных
- Простота и гибкость регулирования гиперпараметров модели

Метод бустинга: AdaBoost (scikit-learn), классификатор Decision Tree

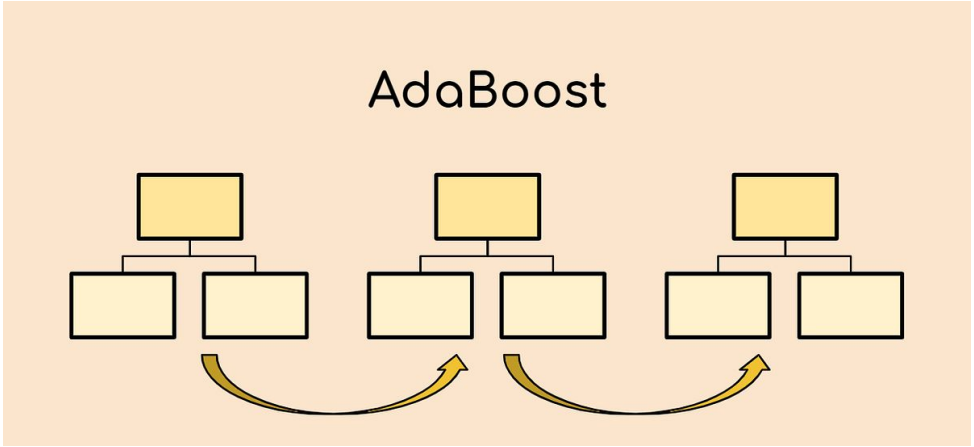


Рис. 8 Схема работы алгоритма AdaBoost

Классификатор на основе BDT

	Importance
Front Time	0.000000
Tail Time	0.172140
FWHM	0.036033
skewness	0.151772
kurtosis	0.157433
Q_total / Q_muon	0.482622

Рис. 9 Оценка BDT важности признаков

Оптимальное количество weak estimator – 20

Accuracy = 0.956 для валидации источником гамма-квантов

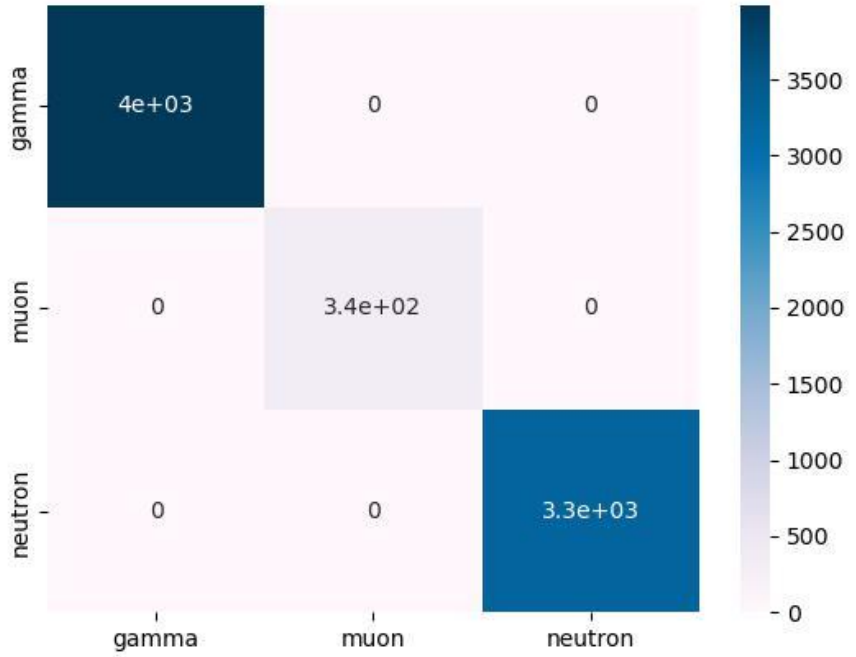


Рис. 10 Confusion-Matrix для модели BDT (по вертикали истинные значения, по горизонтали – предсказания модели)

Классификатор на основе BDT

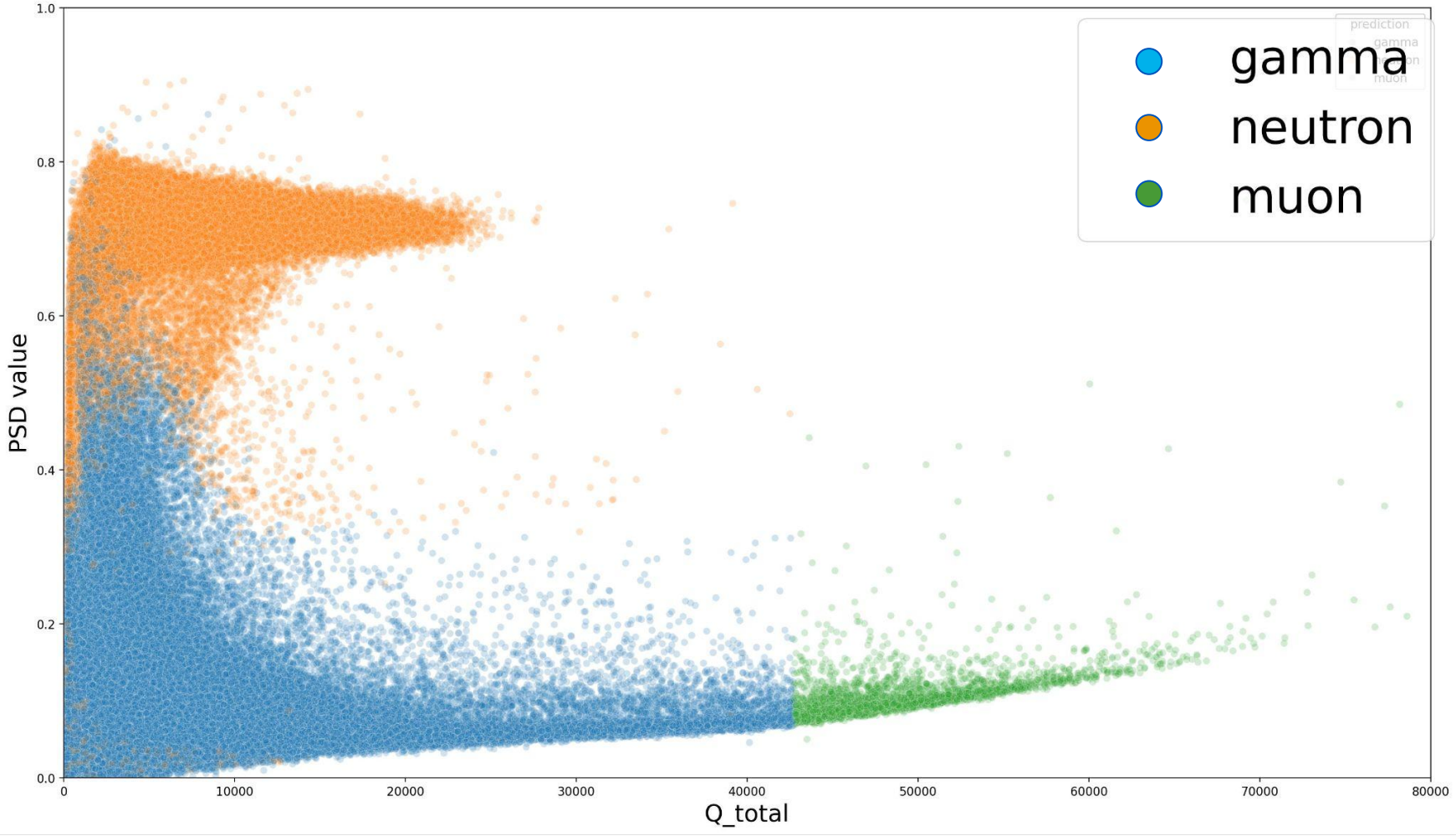


Рис. 11 Разметка данных моделью BDT на графике PSD

Заключение

Таким образом, для эксперимента с событиями типа "нейтрон-гамма" были выполнены:

- Обработка данных и построение признаков для обучения ML-модели
- Обучение и тестирование нескольких классических ML-моделей на экспериментальных данных, изучение характеристик работы моделей
- Валидация работы моделей данными с источником гамма-квантов

Дальнейшие направления развития работы:

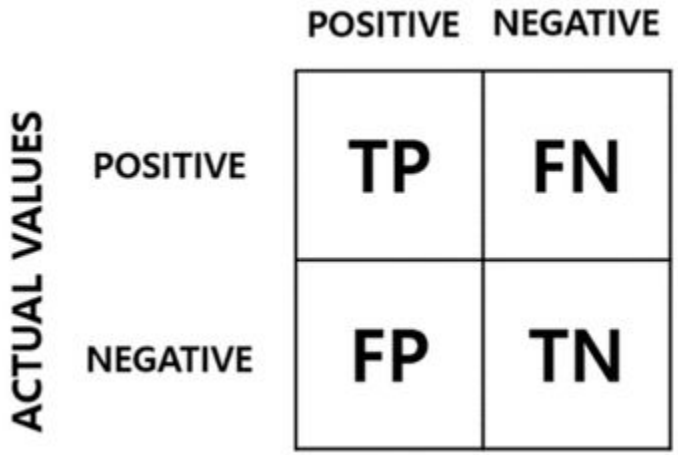
- Применение нейронных сетей для задач классификации частиц
- Создание пайплайна для обработки экспериментальных данных
- Улучшение механизма валидации моделей путём использования моделирования и экспериментальных физических методов
- Реализация методики HANS: модель-генератор сигнала и модель-дискриминатор
- Применение обученных моделей ML для экспериментов с нейтринным детектором

НИЯУ МИФИ

Спасибо за внимание!

Козлов А.А.
andcauselove@yandex.ru

Дополнительный слайд

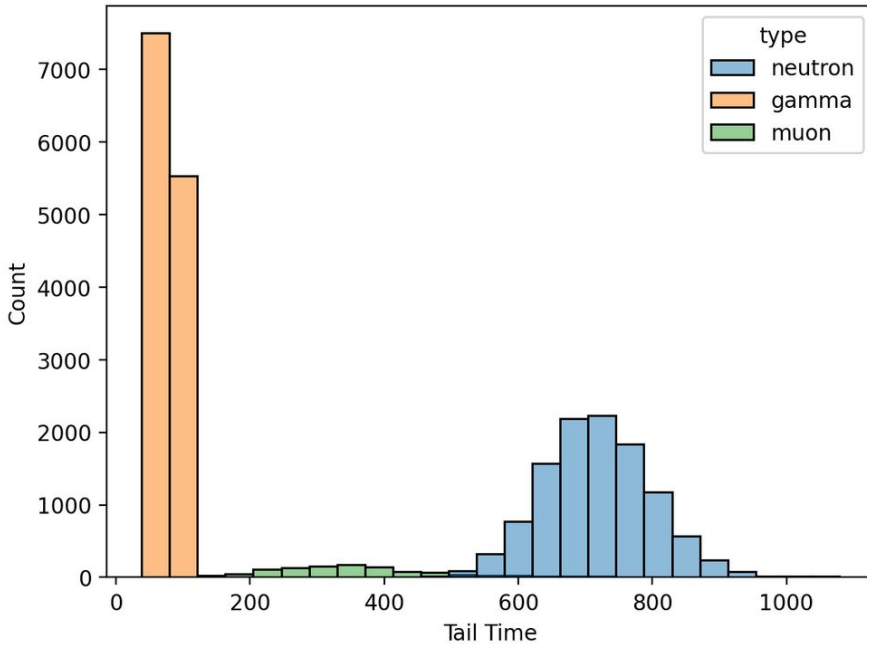


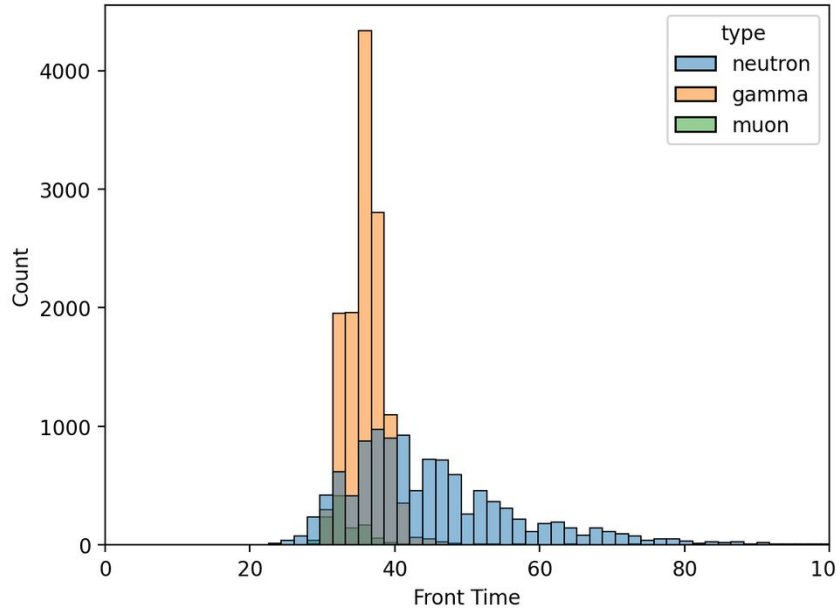
$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

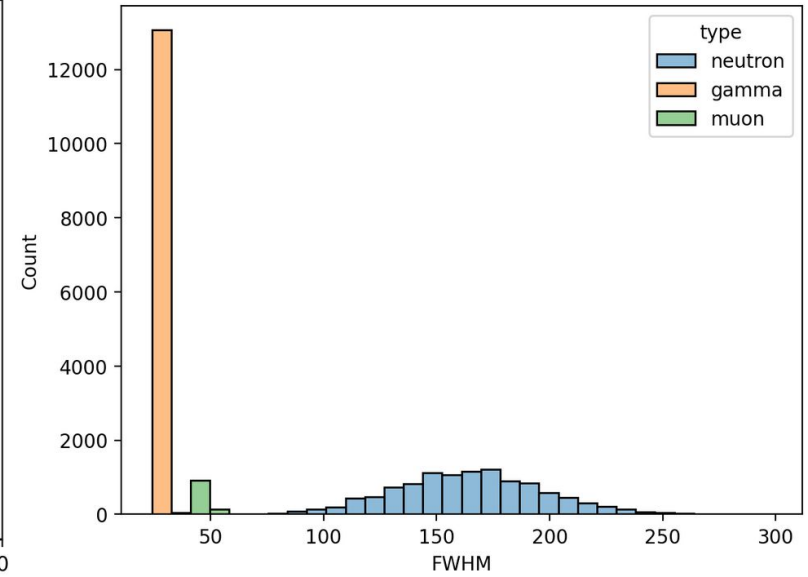
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

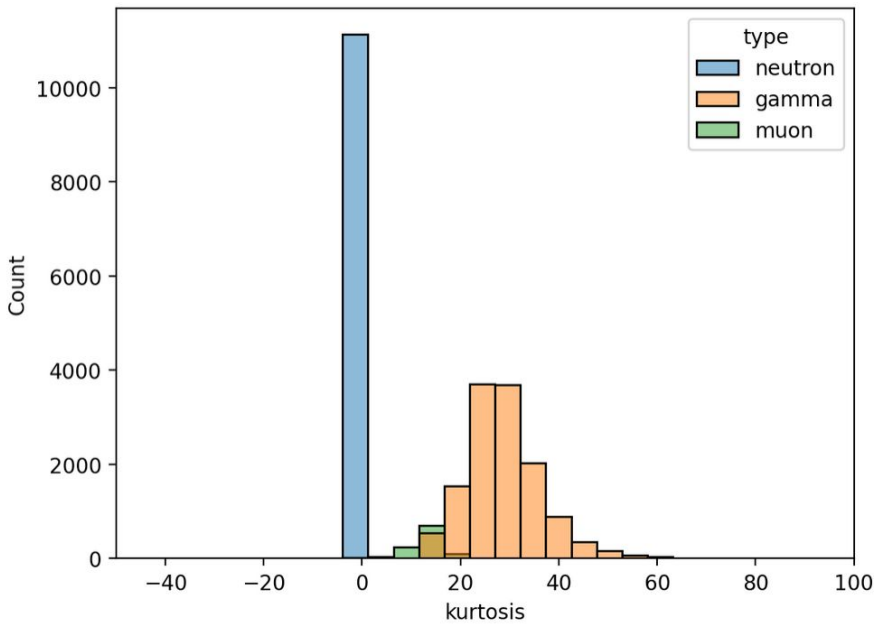
$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

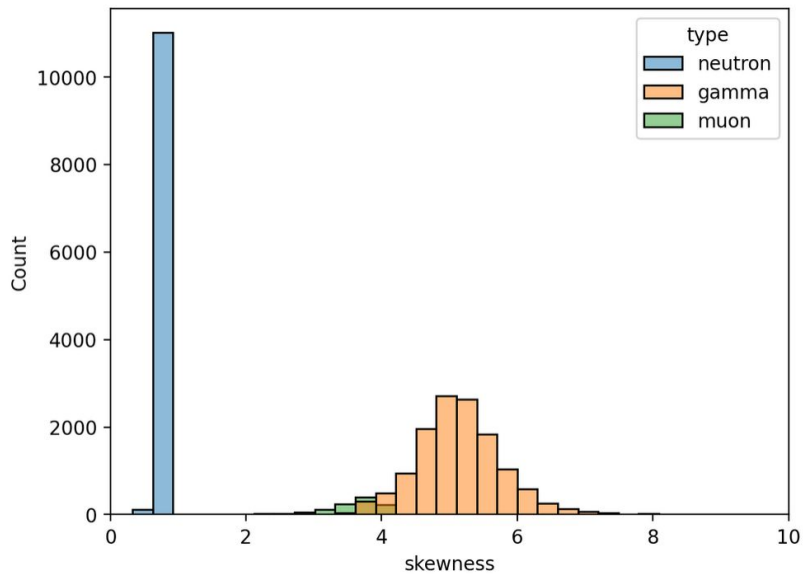
Метрики, использованные для оценки эффективности модели

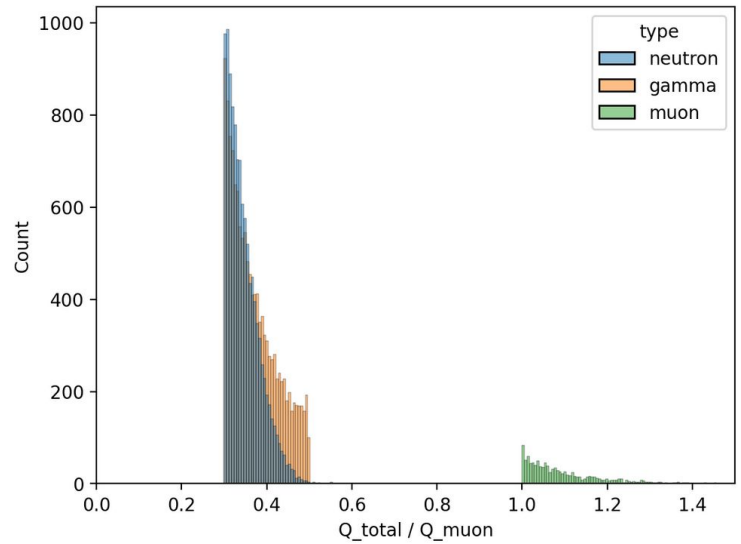

type
neutron
gamma
muon
Count
Tail Time


type
neutron
gamma
muon
Count
Front Time


type
neutron
gamma
muon
Count
FWHM


type
neutron
gamma
muon
Count
kurtosis


type
neutron
gamma
muon
Count
skewness


type
neutron
gamma
muon
Count
Q_total / Q_muon