МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»
(НИЯУ МИФИ)

ИНСТИТУТ ЯДЕРНОЙ ФИЗИКИ И ТЕХНОЛОГИЙ
КАФЕДРА №40 «ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ»

УДК 539.1.072

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА К БАКАЛАВРСКОЙ ДИПЛОМНОЙ РАБОТЕ МОДЕРНИЗАЦИЯ АЛГОРИТМОВ ФОТОННОЙ ИДЕНТИФИКАЦИИ В ЭКСПЕРИМЕНТЕ SPD

Студент _____ Г. Е. Петров

Научный руководитель,
к.ф.-м.н. _____ Е. Ю. Солдатов

Москва 2026

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

МОДЕРНИЗАЦИЯ АЛГОРИТМОВ ФОТОННОЙ ИДЕНТИФИКАЦИИ В ЭКСПЕРИМЕНТЕ SPD

Студент	_____ Г. Е. Петров
Научный руководитель, к.ф.-м.н.	_____ Е. Ю. Солдатов
Рецензент, к.ф.-м.н.	_____ В. С. Воробьев
Секретарь ГЭК, к.ф.-м.н.	_____ А. А. Кириллов
Зав. каф. №40, д.ф.-м.н.	_____ М. Д. Скорохватов

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современные исследования в области физики высоких энергий направлены на изучение фундаментальных свойств материи и взаимодействий на субатомном уровне. Одной из наиболее актуальных задач является исследование структуры нуклона, включая вклад глюонов и кварков в его спин и массу. Для выполнения этой задачи разрабатываются новые экспериментальные установки, такие как Spin Physics Detector (SPD), предназначенный для работы на коллайдере NICA в Объединённом институте ядерных исследований (Дубна).

SPD представляет собой универсальный детектор, созданный для изучения структуры протонов и дейтронов, а также связанных с ними спиновых явлений. Уникальные возможности этой установки позволяют исследовать поляризованные столкновения протонов и дейтронов в ранее недоступных энергетических диапазонах [1], что делает SPD важным дополнением к текущим и будущим экспериментам в таких лабораториях, как BNL, CERN и IHEP CAS.

Одной из основных задач эксперимента SPD является исследование глюонного вклада в спиновую структуру нуклона посредством измерения спиновых асимметрий в столкновениях поляризованных протонов и дейтронов. Для решения этой задачи предполагается использовать несколько взаимодополняющих процессов, чувствительных к глюонным распределениям в нуклоне, включая рождение чармония, открытого чарма и прямых фотонов. Диаграммы, иллюстрирующие данные реакции, представлены на рисунке 1.1.

Среди перечисленных процессов особый интерес представляет образование прямых фотонов. В отличие от адронов, прямые фотоны не участвуют в сильном взаимодействии после рождения и поэтому сохраняют информацию о первичном партонном процессе. Однако их экспериментальное выделение осложняется наличием интенсивного фона от распадов нейтральных

мезонов, прежде всего $\pi^0 \to \gamma\gamma$. При высоких энергиях фотоны от распада π^0 могут формировать в электромагнитном калориметре ливни, практически неотличимые от ливней одиночных фотонов. В связи с этим эффективность физической программы SPD в значительной степени определяется качеством идентификации фотонов и подавления фоновых событий. Настоящая работа посвящена разработке и оптимизации алгоритмов разделения одиночных фотонов и нейтральных π^0-мезонов по характеристикам электромагнитных ливней, регистрируемых электромагнитным калориметром SPD.

Цель работы — разработка универсального инструмента идентификации фотонов в электромагнитном калориметре эксперимента SPD, обеспечивающего возможность применения как методов машинного обучения, так и классических подходов, основанных на фиксированных отборах по физическим наблюдаемым.

Для достижения поставленной цели решались следующие **задачи**:

- исследование архитектуры среды моделирования SPDROOT и текущих алгоритмов реконструкции фотонов в электромагнитном калориметре SPD;
- анализ и адаптация наблюдаемых идентификации фотонов из эксперимента ATLAS применительно к геометрии и конструкции калориметра SPD;
- разработка алгоритмов идентификации фотонов на основе метода boosted decision trees (BDT) и метода фиксированных отборов с использованием адаптированных наблюдаемых;
- ранжирование и отбор наиболее информативных наблюдаемых для кластеров для различных типов классификаторов;
- тестирование разработанных алгоритмов на выборках различной степени реалистичности и сравнение их эффективности с альтернативным MLP-подходом.

Объект исследования — процессы реконструкции и идентификации фотонов в электромагнитном калориметре эксперимента SPD.

Предмет исследования — алгоритмы и наблюдаемые для классификации электромагнитных кластеров для разделения фотонов и нейтральных пионов.

Научная новизна:

- Впервые для геометрии электромагнитного калориметра SPD адаптирован набор наблюдаемых формы ливня, используемых в эксперименте ATLAS, показана их разделяющая способность для задачи γ/π^0 идентификации.
- Впервые в контексте конфигурации электромагнитного калориметра SPD проведён сравнительный анализ трёх подходов к классификации кластеров (BDT, MLP, фиксированные отборы) на основе адаптированных наблюдаемых.
- Впервые для эксперимента SPD выполнен сопоставительный анализ разделяющей способности наблюдаемых из различных экспериментов в контексте идентификации фотонов.

Практическая значимость — разработанный инструмент может быть интегрирован в цепочку реконструкции SPDROOT/SAMPO, обеспечивая гибкий выбор метода идентификации в зависимости от требуемого соотношения эффективности сигнала и режекции фона.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 ЦЕЛЬ ЭКСПЕРИМЕНТА SPD

Одна из основных целей эксперимента SPD - получение доступа к глюонным функциям распределения, зависимым от поперечного импульса (далее будет использоваться аббревиатура TMD PDFs - Transverse Momentum Dependent Parton Distribution Functions) в протоне и дейтроне [2]. Для получения же доступа к TMD PDFs в самом эксперименте будут изучаться такие реакции, как рождение чармония, открытого чарма и прямых фотонов [3]. Диаграммы, иллюстрирующие данные реакции, представлены на рисунке 1.1.

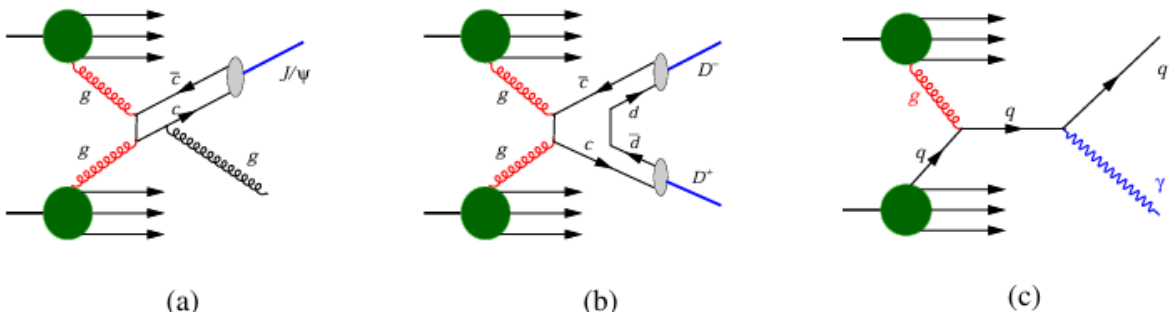

Рисунок 1.1 — Диаграммы, иллюстрирующие реакции: рождение чармония (a), открытого чарма (b) и прямых фотонов (c)

1.2 ПЛАНИРУЕМЫЕ СТАДИИ ЭКСПЕРИМЕНТА И КОНФИГУРАЦИИ УСТАНОВКИ SPD

Программа эксперимента SPD реализуется в два последовательных этапа, различающихся как уровнем доступной светимости и энергий, так и полнотой состава детекторной системы. Такой подход обеспечивает ранний старт физической программы при ограниченной инфраструктуре и последующее расширение установки до полной конфигурации, необходимой для измерений глюонной структуры нуклона.

Первый этап: базовая конфигурация установки

Первый этап ориентирован на проведение исследований при пониженной светимости и энергии столкновений с использованием поляризованных протонных и дейтронных пучков, а также ионных столкновений. Основной акцент делается на измерении спин-зависимых эффектов в эксклюзивных и полуинклюзивных процессах, включая упругое рассеяние, рождение гиперонов, дибарионных резонансов и околопороговое образование чармония [4].

На данном этапе планируется использовать: сверхпроводящий соленоид (Magnet System), трекер на дрейфовых трубках (ST), микромегас центральный трекер (MCT), «пробежную» систему (Range System, RS), систему счетчиков ”пучок-пучок”(Beam-Beam Counters, BBC), нуль градусные калориметры (Zero Degree Calorimeters, ZDC) [4].

Конфигурация детекторов первого этапа показана на рисунке 1.2.

Сверхпроводящий соленоид (Magnet System) Соленоид обеспечивает однородное магнитное поле величиной до 1 Тл в области взаимодействия. Поле используется для измерения импульсов заряженных частиц по кривизне треков. Конструктивно магнит выполнен в виде сверхпроводящей катушки, размещённой внутри стального ярма, которое одновременно выполняет функцию поглотителя и элемента возврата магнитного потока. Совместно с трековыми системами магнит обеспечивает относительное разрешение по импульсу порядка 1–2% при $p \approx 1$ ГэВ/с [4].

Трекер на дрейфовых трубках (Straw Tracker, ST) Трекер основан на тонкостенных дрейфовых трубках (straw tubes), заполненных газовой смесью

Ar/CO_2. Заряженная частица ионизирует газ, а дрейф электронов к аноду позволяет восстановить координату прохождения частицы. Геометрия включает баррельную и эндкап-части, обеспечивая почти 4π покрытие. Типичное пространственное разрешение составляет порядка 100–150 мкм. Помимо трекинга, измерение удельных потерь энергии (dE/dx) используется для дополнительной идентификации частиц на низких импульсах [4].

Микромегас-центральный трекер (MCT) На первом этапе используется как компактный внутренний трекер, компенсирующий отсутствие кремниевого вершинного детектора. Детектор основан на технологии Micromegas — газовом усилителе с микросеткой (micromesh), разделяющей дрейфовый и усилительный зазоры. Первичная ионизация в дрейфовом объёме приводит к лавинному усилению электронов в тонком (≈120 мкм) зазоре. MCT улучшает реконструкцию импульсов и эффективность трекинга в центральной области, но не обеспечивает полноценного восстановления вторичных вершин [4].

Range System (RS) Range System представляет собой многослойную систему на основе мини-дрейфовых трубок, использующую стальные элементы ярма в качестве поглотителя. Основная задача — идентификация мюонов и подавление адронного фона. Мюоны, в отличие от адронов, проникают глубже в слои системы, что позволяет их эффективно отделять по характеру пробега. Также RS обеспечивает грубую адронную калориметрию и частично регистрирует нейтроны [4].

Beam-Beam Counters (BBC) BBC представляют собой сцинтилляционные детекторы с сегментированными секторами, расположенные в области больших псевдобыстрот. Основные задачи: локальная поляриметрия, определение плоскости реакции, мониторинг светимости и оценка времени взаимодействия t_0. Принцип работы основан на регистрации заряженных частиц через сцинтилляцию и последующее считывание сигнала SiPM. Азимутальные асимметрии в распределении частиц используются для извлечения информации о поляризации пучка [4].

Zero Degree Calorimeters (ZDC) ZDC расположены вблизи оси пучка за отклоняющими магнитами и предназначены для регистрации нейтронов и фотонов, летящих в нулевой угол. Конструкция основана на чередовании сцинтилляционных плит и вольфрамовых поглотителей с SiPM-считыванием.

Основные задачи: измерение светимости, локальная поляриметрия по нейтронам-спектаторам и временная привязка событий. Высокая радиационная стойкость и временное разрешение порядка 150–200 пс обеспечивают работу в условиях высокой интенсивности пучка [4].

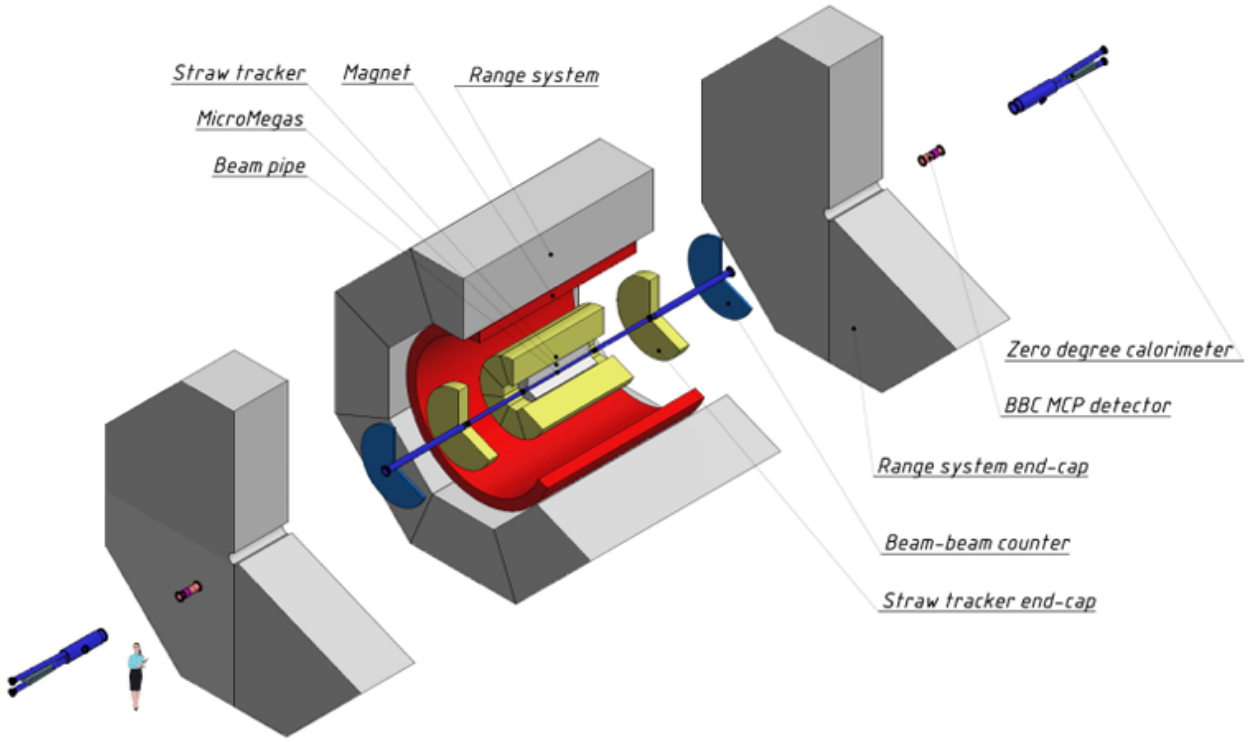

Рисунок 1.2 — Детекторы первой стадии эксперимента

Второй этап: полная конфигурация установки

Второй этап соответствует полной реализации SPD и ориентирован на основную физическую программу — изучение глюонной структуры нуклона через измерение спин-зависимых асимметрий в процессах рождения прямых фотонов, открытого чарма и чармония при максимальных энергиях и светимости.

На данном этапе дополнительно планируется использовать: кремниевый вершинный детектор (SVD), электромагнитный калориметр (ECal), система времени пролёта (TOF), фокусирующий аэрогельный черенковский детектор (FARICH) [4].

Полная конфигурация установки показана на рисунке 1.3.

Кремниевый вершинный детектор (SVD) SVD состоит из нескольких цилиндрических слоёв, расположенных максимально близко к точке вза-

имодействия. Основная задача — точное восстановление первичных и вторичных вершин, что критично для идентификации D-мезонов и других короткоживущих частиц. Используются MAPS-сенсоры, в которых чувствительный слой и электроника интегрированы в одном кристалле. Это обеспечивает низкий шум, высокую гранулярность и радиационную стойкость. Пространственное разрешение достигает порядка нескольких микрометров [4].

Электромагнитный калориметр (ECal) ECal предназначен для регистрации фотонов и заряженных частиц (в основном электронов) в диапазоне энергий от сотен МэВ до 10 ГэВ. Конструкция основана на sampling-структуре (сцинтиллятор + свинец) с высокой сегментацией. При прохождении частицы образуется электромагнитный ливень, энергия которого пропорциональна суммарному световому сигналу. Энергетическое разрешение порядка $5\%/\sqrt{E}\oplus 1\%$ обеспечивает выделение прямых фотонов на фоне распадов π^0 [4].

Система времени пролёта (TOF) TOF используется для идентификации заряженных частиц по времени пролёта на фиксированном расстоянии от точки взаимодействия. Основана на технологии MRPC (Multigap Resistive Plate Chambers), где сигнал формируется лавиной в многозазорной структуре газового детектора. Временное разрешение 50–70 пс позволяет эффективно разделять π/K/p до импульсов порядка нескольких ГэВ/c. TOF также участвует в определении времени события t_0 и синхронизации треков [4].

FARICH (Focusing Aerogel RICH) FARICH обеспечивает идентификацию частиц в промежуточном импульсном диапазоне (0.6–5 ГэВ/c) через регистрацию черенковского излучения в аэрогеле с переменным показателем преломления. Многослойная структура аэрогеля формирует сфокусированные или мультикольцевые изображения, уменьшая неопределённость измерения угла. Фотоны регистрируются матрицей фотодетекторов, а реконструкция кольца позволяет разделять π/K с уровнем 2–3σ [4].

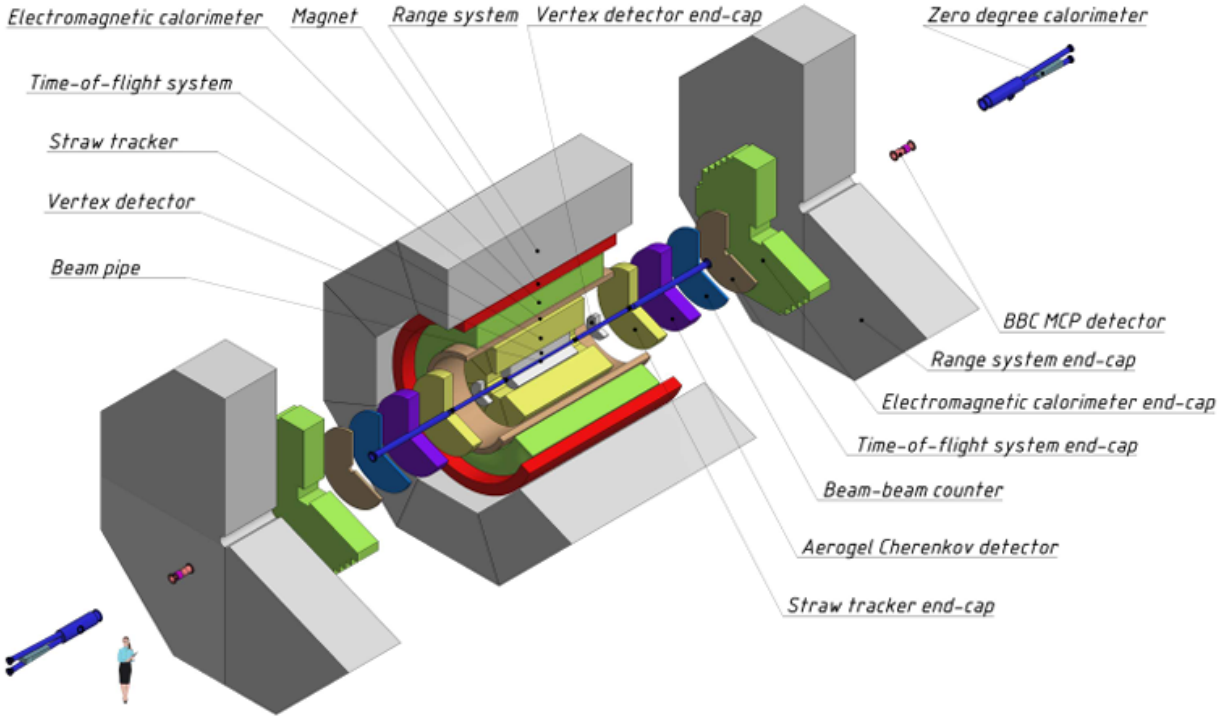

Рисунок 1.3 — Детекторы второй стадии эксперимента

1.3 НАЗНАЧЕНИЕ И КОНСТРУКЦИЯ ЭЛЕКТРОМАГНИТНОГО КАЛОРИМЕТРА SPD

Электромагнитный калориметр (ECal) является одним из ключевых компонентов детектора SPD и предназначен для точного измерения энергии, координат и времени прихода фотонов и заряженных частиц (в основном электронов), возникающих в результате столкновений частиц.

Конструктивно калориметр представляет собой сэмплирующий детектор, состоящий из чередующихся слоёв свинца (поглотитель) и пластикового сцинтиллятора. Каждый модуль включает 190 двойных слоев из 1.5 мм сцинтиллятора и 0.5 мм свинца, что обеспечивает суммарную толщину активной части порядка 380 мм. Свет от сцинтилляторов собирается с помощью волокон со сдвигом длины волны (wavelength-shifting fibers, или WLS) и регистрируется многоэлементными фотодиодами (multi-pixel photon counter, или MPPC) [4]. Общая толщина модуля с учётом конструктивных элементов составляет около 490 мм. Чертёж отдельного модуля калориметра представлен

на рисунке 1.4.

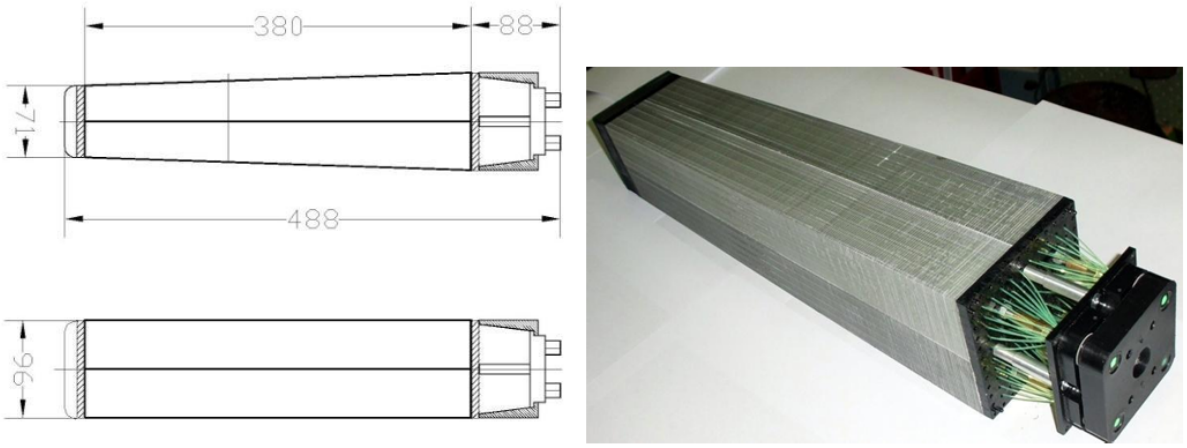

Рисунок 1.4 — Чертеж отдельного модуля калориметра (a) и фото модуля без внешнего корпуса (b)

Калориметр разделён на цилиндрическую (баррельную или центральную) часть и два торцевых эндкапа, что позволяет покрывать почти полный 4π телесный угол: по ϕ от 0 до 2π, по η приблизительно от -3.2 до 3.2. Общий вид калориметра показан на рисунке 1.5.

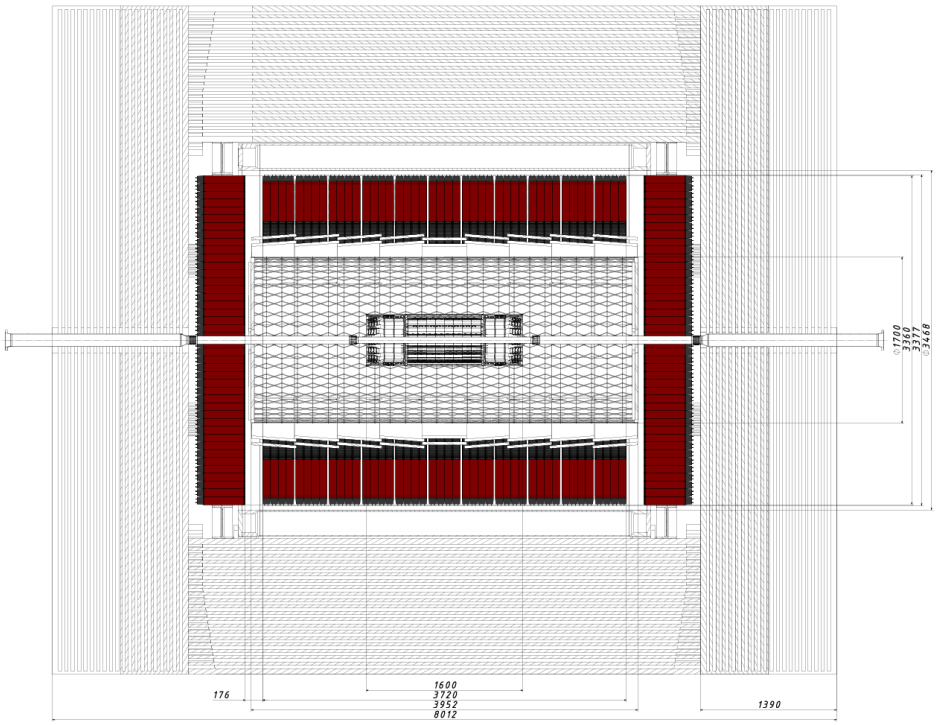

Рисунок 1.5 — Калориметр, вид сбоку, красным изображены эндкапы и цилиндрическая часть калориметра

2 АНАЛИЗ СУЩЕСТВУЮЩИХ АЛГОРИТМОВ И МЕТОДИКА ИССЛЕДОВАНИЯ

2.1 ОБЩЕЕ ОПИСАНИЕ СРЕДЫ SPDROOT

SPDROOT является специализированной средой моделирования и анализа данных, разработанной для эксперимента Spin Physics Detector (SPD) на коллайдере NICA [5]. Эта среда как свою базу использует инструментарий пакета ROOT, который широко используется в экспериментах в области физики высоких энергий для обработки и анализа больших объемов научных данных. SPDROOT предоставляет инструменты для моделирования детекторов, генерации событий, реконструкции частиц и анализа данных. SPDROOT включает в себя модули для симуляции взаимодействий частиц с детекторами, что позволяет исследователям оценивать эффективность различных алгоритмов реконструкции и оптимизировать параметры детекторов.

На момент написания данной работы в среде SPDROOT реализованы базовые алгоритмы кластеризации и проведены первые исследования по разделению кластеров, соответствующих фотонам, от кластеров, соответствующих π^0 [6].

2.2 ТЕКУЩИЕ АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

Кластер формируется следующим образом: Процесс повторяется до тех пор, пока не добавлены все хиты.ааллфя берётся любой хит (срабатывание

ячейки - минимального детектирующего элемента электромагнитного калориметра) с энергией выше порога (0.02 ГэВ), объявляется началом кластера, после чего в этот кластер добавляются все остальные хиты, которые находятся достаточно близко (минимальное расстояние 10 см) хотя бы к одному уже добавленному хиту. Процесс повторяется до тех пор, пока не добавлены все хиты.

2.3 ТЕКУЩИЕ АЛГОРИТМЫ ИДЕНТИФИКАЦИИ И ПОСТАНОВКА ЗАДАЧИ

Для кластеров от π^0 существует 2 принципиальных случая (другие типы кластеров маловероятны либо будут идентифицироваться другими детекторами):

1) Фотоны от π^0 попали в один кластер (условное обозначение $[\pi^0, \gamma, \gamma]$): тогда такой кластер имеет специфическую форму
2) Фотоны от π^0 попали ва разные кластеры (условное обозначение $[\pi^0, \gamma]$): такие кластеры можно идентифицировать из кинематики

Кластеры первого типа представляют особый интерес, поскольку их идентификация невозможна исключительно на основе кинематических соображений и требует анализа формы энерговыделения в ячейках калориметра.

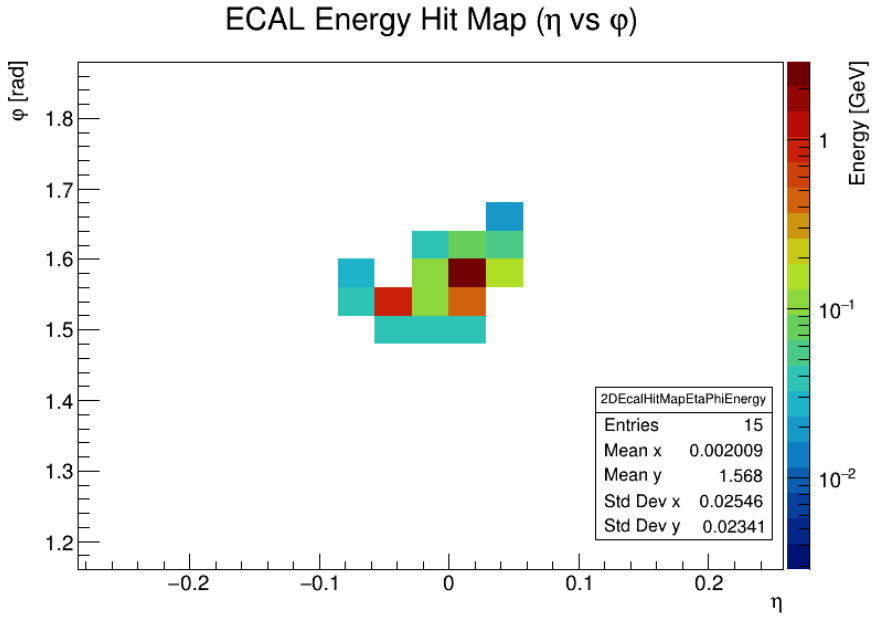

Рисунок 2.1 — Пример кластера частицы π^0, запущенной с энергией 5 ГэВ

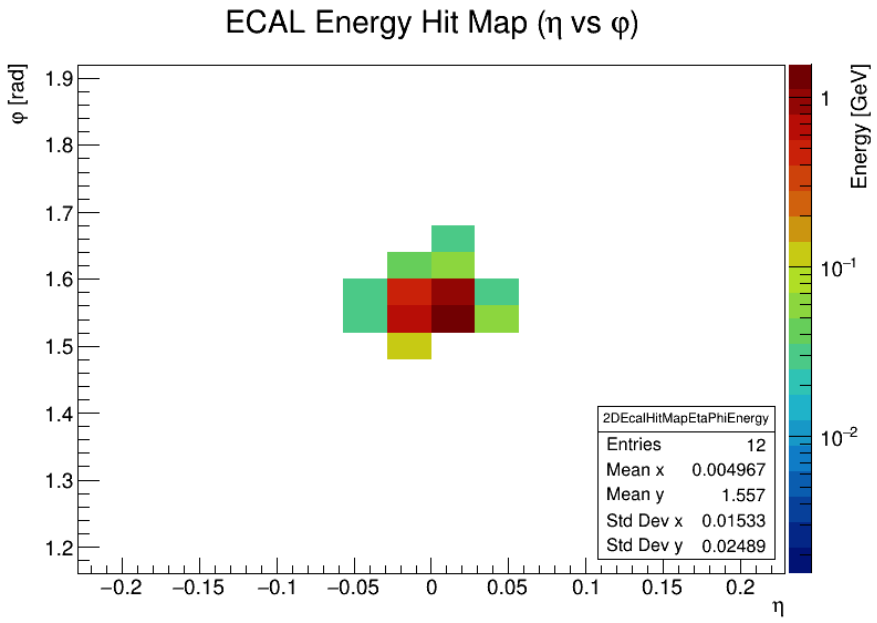

Рисунок 2.2 — Пример кластера частицы γ, запущенной с энергией 5 ГэВ

В текущих алгоритмах идентификации, разработанных до этой работы, решается задача разделения именно кластеров 1го типа (пример на рисунке 2.1) и кластеров от прямых фотонов (пример на рисунке 2.2). Разделение кластеров фотонов и π^0 проводится на основе наблюдаемых, зависящих от формы распределения энергии внутри кластера и основанных на методологии эксперимента LHCb [6]. Подробное описание наблюдаемых будет приведено в секции 3.2.3.

Далее используется MLP нейронная сеть для классификации кластеров на фотонные и π^0-кластеры. В текущей версии алгоритмов разделения были достигнуты следующие результаты: при 80% γ эффективности около 90% π^0 режекция [7]. Данный подход является лишь предварительной попыткой разделения и данная задача требует более комплексного анализа.

В данной работе планируется систематизировать работу по идентификации фотонов в SPD: использовать разные подходы и разработать инструмент для идентификации с заданной эффективностью сигнала и максимально возможной режекцией фона.

2.4 ОПИСАНИЕ ПОДХОДОВ К КЛАССИФИКАЦИИ КЛАСТЕРОВ

В данной работе были реализованы и протестированы несколько подходов к классификации кластеров в электромагнитном калориметре SPD с целью разделения фотонных и π^0-кластеров. Сравнивались следующие методы:

- **Метод фиксированных отборов (Rectangular Cuts).** Данный метод основан на применении прямоугольных пороговых отсечек по набору физических наблюдаемых для кластера. Оптимальные значения порогов подбираются автоматически в TMVA (Toolkit for Multivariate Data Analysis with ROOT [8]) путём перебора допустимых диапазонов наблюдаемых с целью максимизации режекции фона при заданной эффективности. Метод является интерпретируемым, но не учитывает корреляции между наблюдаемыми. Для выбора наилучшего набора порогов используется метрика *EffSel* — эффективность селекции сигнала на тренировочной выборке, которая максимизирует выделение сигнала при заданном уровне подавления фона [8].
- **MLP (Multi-Layer Perceptron).** Многослойный персептрон представляет собой нейросетевой классификатор, способный моделировать нелинейные зависимости между входными наблюдаемыми [9]. Сеть состоит из входного слоя, одного или нескольких скрытых слоев с нелинейными активационными функциями и выходного слоя, который выдаёт вероятность принадлежности к классу. (данный метод уже был реализован в текущей версии алгоритмов идентификации кластеров в SPDROOT[7]).
- **BDT (Boosted Decision Trees).** Метод основан на ансамбле деревьев решений, обучаемых с использованием бустинга (AdaBoost, Gradient Boosting) [8]. Каждое дерево выполняет последовательные бинарные разбиения пространства признаков, а итоговое решение формируется как взвешенная сумма ответов отдельных деревьев. BDT эффективно учитывает корреляции между наблюдаемыми, устойчив к выбросам и, как правило, демонстрирует высокую разделяющую способность при анализе многомерных данных. Для подбора оптимального разбиения в

каждом узле используется *Gini Index*:

$$\text{Gini} = 2 \cdot p_s \cdot (1 - p_s),$$

где p_s — доля сигналов в узле. Деревья строятся так, чтобы суммарно максимизировать разделение сигнал/фон по всей тренировочной выборке [8].

2.5 МЕТОДИКА ОЦЕНКИ КАЧЕСТВА КЛАССИФИКАЦИИ

В секции 3.2 в качестве сигнала понимались все кластеры от фотонов, а в качестве фона - все кластеры от π^0 (за исключением отборов описанных в секции). Сама по себе секция 3.2 имеет промежуточный характер, в которой описан сам процесс исследования, поэтому результаты в ней полученные, не относятся к конечным, а понимание сигнала/фона отличается.

В секции 3.3 рассматривались 2 основных сценария разделения сигнал/фон (в зависимости от подтипа кластера - подробнее в начале секции 3.3).

Для оценки качества классификации в данной работе используются следующие метрики.

Эффективность сигнала (ε_S) определяется как доля сигнальных событий, правильно классифицированных как сигнал:

$$\varepsilon_S = \frac{N_{S \to S}}{N_S},$$

где N_S — общее количество сигнальных событий, $N_{S \to S}$ — количество сигнальных событий, правильно идентифицированных.

Режекция фона (R_B) определяется как доля фоновых событий, правильно отбракованных классификатором:

$$R_B = 1 - \frac{N_{B \to S}}{N_B},$$

где N_B — общее количество фоновых событий, $N_{B\to S}$ — количество фоновых событий, ошибочно классифицированных как сигнал.

В задачах идентификации частиц в физике высоких энергий часто используется ROC-кривая (Receiver Operating Characteristic, рабочая характеристика приемника), отражающая зависимость эффективности сигнала от эффективности (или режекции) фона при варьировании порога классификации.

В данной работе под ROC-кривой понимается зависимость режекции фона от эффективности сигнала — данное представление наиболее наглядно в контексте решаемой задачи, хотя и не является классическим определением. Разделяющая способность классификатора тем выше, чем ближе ROC-кривая расположена к правому верхнему углу графика. Примеры ROC-кривых представлены на рисунке 2.3.

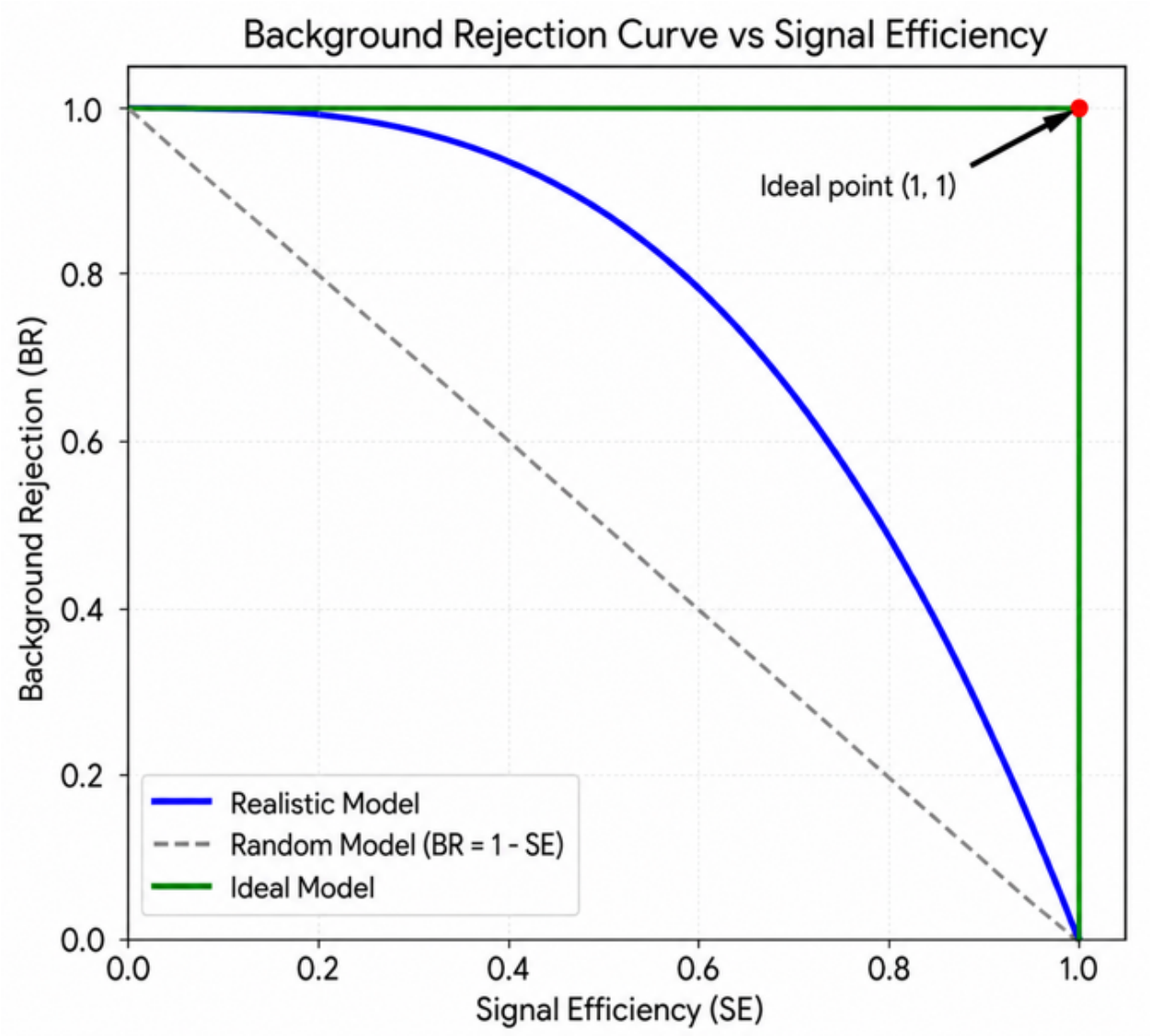

Рисунок 2.3 — Примеры ROC-кривых: пунктиром - случайная модель, синим - произвольная модель, зеленым - идеальная модель

В качестве основной количественной метрики качества классификации в данной работе выбрана площадь под ROC-кривой (ROC AUC — Area Under Curve). Данная метрика является общепринятым показателем качества бинарной классификации и позволяет проводить объективное сравнение различных алгоритмов независимо от выбора порога классификации. Значение ROC AUC, равное 1, соответствует идеальному разделению, значение 0.5 — полному отсутствию разделяющей способности. Выбор ROC AUC в качестве основной метрики обусловлен следующими преимуществами:

- возможность сравнения классификаторов при различных порогах срабатывания без привязки к конкретному порогу;

- устойчивость к дисбалансу классов (сигнал и фон) в выборке;
- широкое распространение в задачах идентификации частиц в физике высоких энергий, что обеспечивает сопоставимость результатов.

3 РАЗРАБОТКА АЛГОРИТМОВ ИДЕНТИФИКАЦИИ ЧАСТИЦ В ЭЛЕКТРОМАГНИТНОМ КАЛОРИМЕТРЕ SPD

3.1 АНАЛИЗ НАБЛЮДАЕМЫХ, ИСПОЛЬЗУЕМЫХ В ЭКСПЕРИМЕНТЕ ATLAS, НА ПРИМЕНИМОСТЬ К ИДЕНТИФИКАЦИИ ЧАСТИЦ В ЭЛЕКТРОМАГНИТНОМ КАЛОРИМЕТРЕ SPD

Задача фотонной идентификации (а именно выделение фотонов среди подавляющего адронного фона) уже решалась во многих экспериментах на адронных коллайдерах, например в эксперименте ATLAS. В эксперименте ATLAS были использованы следующие наблюдаемые [10]:

Покрытие детектора. Кандидаты в фотоны должны удовлетворять условию

$$|\eta| < 2.37,$$

за исключением переходной области калориметра

$$1.37 < |\eta| < 1.52.$$

Адронная утечка. Доля поперечной энергии, утекающей в адронный калориметр, используется для подавления адронного фона. В зависимости от псевдобыстроты определены две наблюдаемые:

- R_{had1}: отношение поперечной энергии, выделенной в первом слое адронного калориметра, к поперечной энергии электромагнитно-

го кластера; используется в диапазонах

$$|\eta| < 0.8 \quad \text{и} \quad |\eta| > 1.37;$$

- R_{had}: отношение поперечной энергии, выделенной во всём адронном калориметре, к поперечной энергии электромагнитного кластера; используется в диапазоне

$$0.8 < |\eta| < 1.37.$$

Наблюдаемые второго слоя калориметра. Наблюдаемые, характеризующие поперечное развитие электромагнитного ливня во втором слое калориметра:

- R_η: отношение энергии, выделенной в окне 3×7 ячеек, к энергии в окне 7×7 ячеек в направлении η;
- w_2: поперечная ширина электромагнитного ливня во втором слое калориметра;
- R_ϕ: отношение энергии, выделенной в окне 3×3 ячеек, к энергии в окне 3×7 ячеек в направлении ϕ.

Наблюдаемые стрипового слоя калориметра. Наблюдаемые, использующие тонкую сегментацию первого (стрипового) слоя калориметра и особенно чувствительные к распадам $\pi^0 \to \gamma\gamma$:

- w_{s3}: ширина ливня, вычисленная по трём стрипам, центрированным на стрипе с максимальной энергией;
- w_{stot}: полная поперечная ширина ливня в стриповом слое;
- F_{side}: доля энергии, выделенной вне ядра трёх центральных стрипов, но внутри окна из семи стрипов;
- ΔE: разность между энергией второго локального максимума в стриповом слое и минимальной энергией, реконструированной в стрипе между первым и вторым максимумами;
- E_{ratio}: отношение

$$E_{\text{ratio}} = \frac{E_{\text{max1}} - E_{\text{max2}}}{E_{\text{max1}} + E_{\text{max2}}},$$

где E_{max1} и E_{max2} — наибольшее и второе по величине энерговы-

деления в стриповом слое.

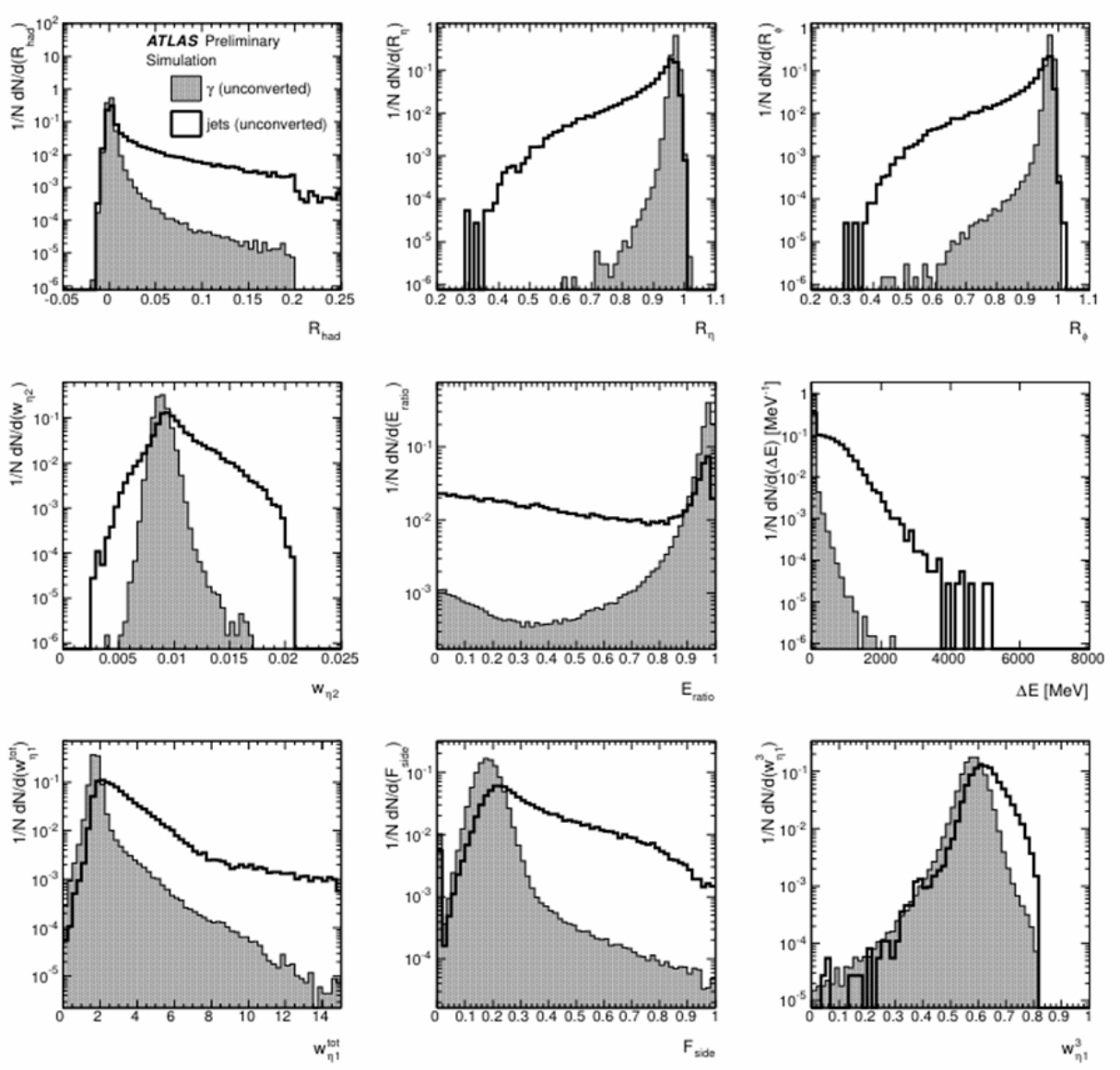

Рисунок 3.1 — Распределения наблюдаемых, зависящих от формы кластера, в эксперименте ATLAS

Поскольку электромагнитный калориметр SPD имеет отличную от ATLAS конструкцию и геометрию, не все перечисленные наблюдаемые могут быть применены напрямую. Среди данных наблюдаемых были выбраны те, которые можно адаптировать к модели электромагнитного калориметра SPD. Были реализованы функции для вычисления следующих адаптированных наблюдаемых для кластеров:

- R_η: отношение энергии, выделенной в окне 3×7 ячеек, к энергии в

окне 7×7 ячеек по (η, ϕ);

- $w_{\eta 2}$: поперечная ширина электромагнитного ливня в ячейках кластера

$$w_{\eta 2} = \sqrt{\frac{\sum_i E_i \eta_i^2}{\sum_i E_i} - \left(\frac{\sum_i E_i \eta_i}{\sum_i E_i}\right)^2}$$

сумма по окну 3×5 ячеек по (η, ϕ);

- R_ϕ: отношение энергии, выделенной в окне 3×3 ячеек, к энергии в окне 3×7 ячеек по (η, ϕ);
- E_{ratio}: отношение

$$E_{\text{ratio}} = \frac{E_{\text{max1}} - E_{\text{max2}}}{E_{\text{max1}} + E_{\text{max2}}},$$

где E_{max1} и E_{max2} — наибольшее и второе по величине энерговыделения в ячейках калориметра;

- ΔE: разность между энергией второго локального максимума в ячейках кластера и минимальной энергией, реконструированной в прямоугольнике между первым и вторым максимумами. В случае отсутствия второго максимума наблюдаемая принимает значение 0.

Для оценки эффективности адаптированных наблюдаемых были проведены тестовые запуски фотонов и π^0 в среде SPDROOT с последующим вычислением адаптированных наблюдаемых для каждого кластера. Запуск производился с фиксированной энергией 4 ГэВ перпендикулярно центральной части калориметра. На данном этапе отбор среди кластеров π^0 подтипа кластеров [π^0, γ, γ] не производился. Были построены гистограммы распределений адаптированных наблюдаемых для кластеров, соответствующих фотонам и π^0. Данные гистограммы изображены на рисунках 3.2, 3.3, 3.4, 3.5 и 3.6.

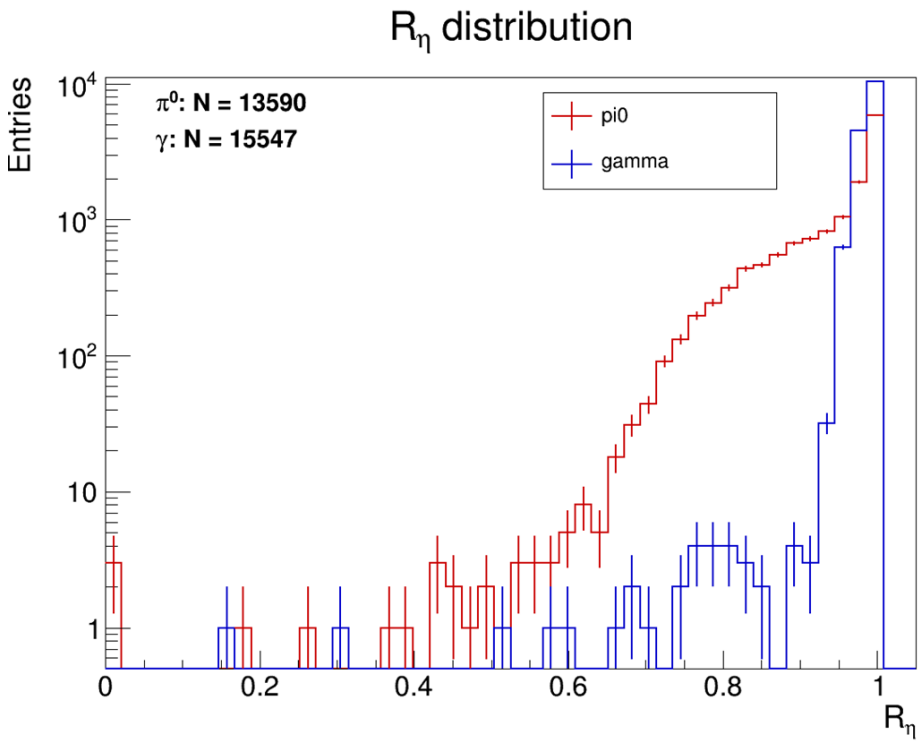

Рисунок 3.2 — Распределение адаптированной наблюдаемой R_η, 10000 событий γ (синим) и π^0 (красным), запущенных с энергией 4 ГэВ

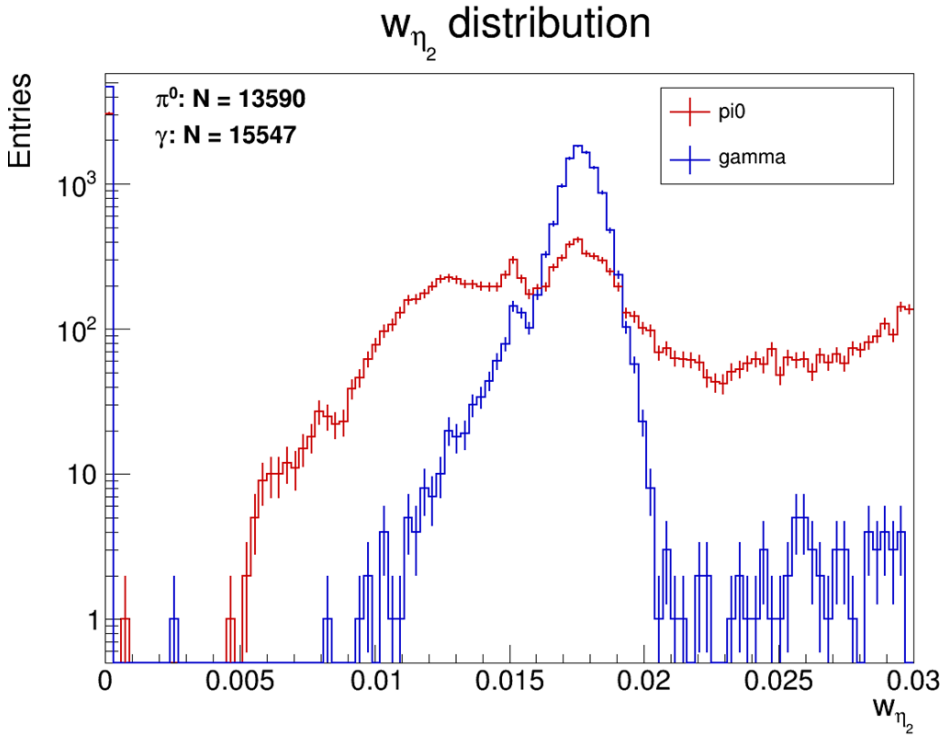

Рисунок 3.3 — Распределение адаптированной наблюдаемой $w_{\eta 2}$, 10000 событий γ (синим) и π^0 (красным), запущенных с энергией 4 ГэВ

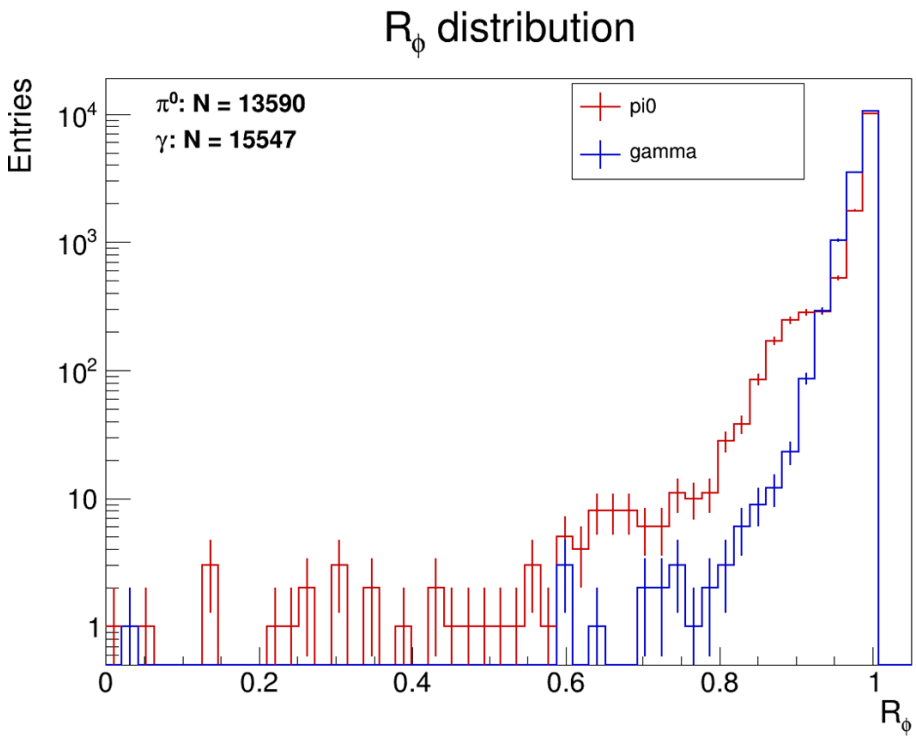

Рисунок 3.4 — Распределение адаптированной наблюдаемой R_ϕ, 10000 событий γ (синим) и π^0 (красным), запущенных с энергией 4 ГэВ

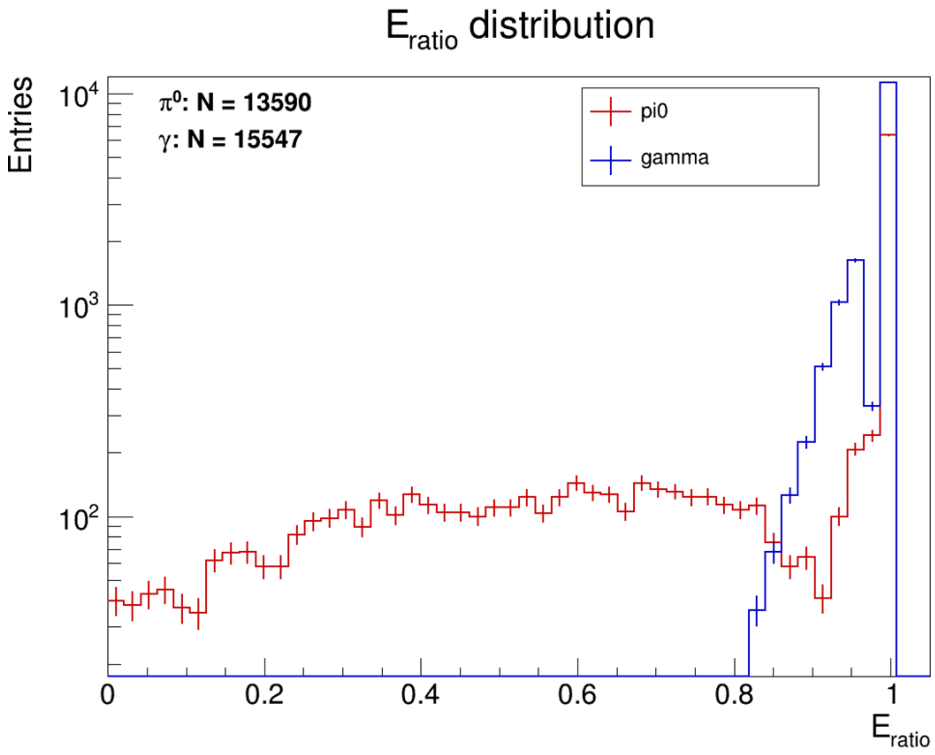

Рисунок 3.5 — Распределение адаптированной наблюдаемой E_{ratio}, 10000 событий γ (синим) и π^0 (красным), запущенных с энергией 4 ГэВ

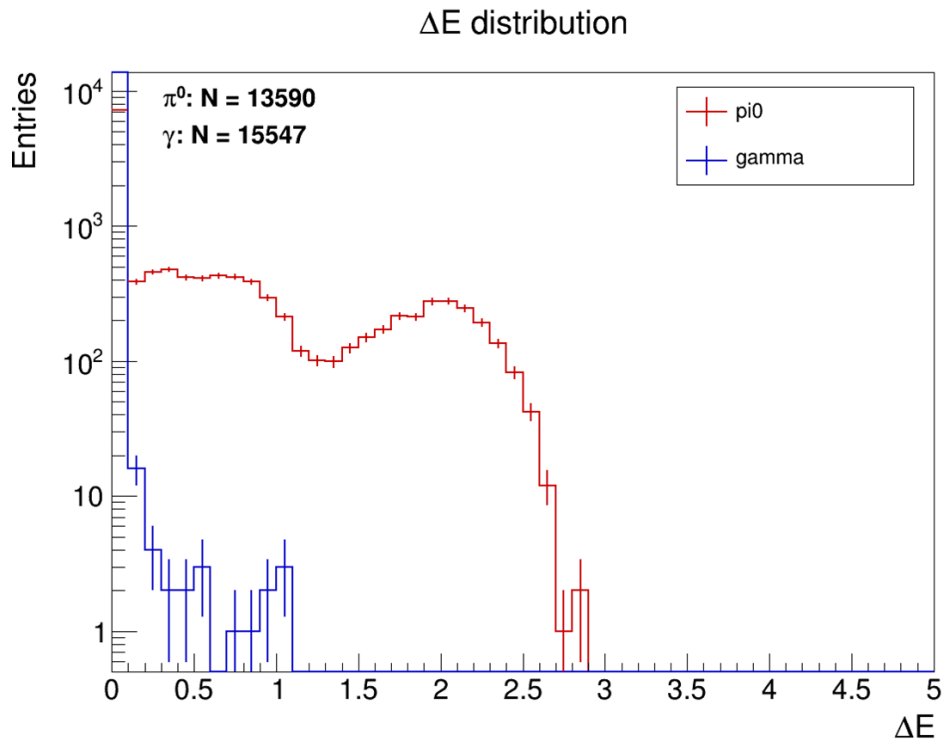

Рисунок 3.6 — Распределение адаптированной наблюдаемой ΔE, 10000 событий γ (синим) и π^0 (красным), запущенных с энергией 4 ГэВ

Анализ полученных распределений показал, что новые наблюдаемые демонстрируют различия между фотонными и π^0-кластерами, что свидетельствует об их потенциальной полезности для задачи разделения. Количественная оценка их ”полезности”по различным метрикам для конкретных случаев будет приведена позднее в данной работе.

3.2 СРАВНЕНИЕ ЭФФЕКТИВНОСТИ РАЗЛИЧНЫХ АЛГОРИТМОВ КЛАССИФИКАЦИИ КЛАСТЕРОВ НА ПРОСТЫХ ВЫБОРКАХ

3.2.1 ОБУЧЕНИЕ И ТЕСТИРОВАНИЕ BDT КЛАССИФИКАТОРА С ИСПОЛЬЗОВАНИЕМ АДАПТИРОВАННЫХ НАБЛЮДАЕМЫХ ДЛЯ КЛАСТЕРОВ ИЗ ЭКСПЕРИМЕНТА ATLAS

Для обучения BDT классификатора была подготовлена выборка из 10000 событий γ и 10000 событий π^0, запущенных с энергиями 2, 4, 8 ГэВ перпендикулярно цилиндрической части калориметра. Отбор на подтип кластера на данном этапе не производился. 70% выборки было использовано для обучения, 30% — для тестирования.

Были построены корреляционные матрицы для адаптированных наблюдаемых для кластеров. Матрицы показали наличие корреляций между некоторыми из наблюдаемых. Но в целом корреляции были не слишком сильными (в частности некоторые наблюдаемые, которые коррелировали для одного типа кластеров, не коррелировали для другого), поэтому на этом начальном этапе были выбраны все наблюдаемые (из ранее описанных) в качестве входных для BDT классификатора.

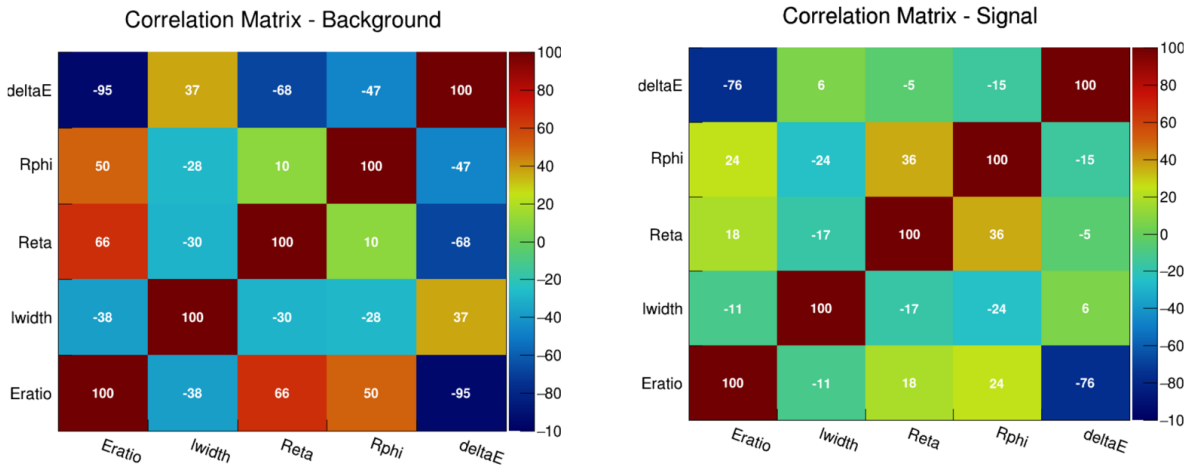

Рисунок 3.7 — Корреляционные матрицы адаптированных наблюдаемых для кластеров

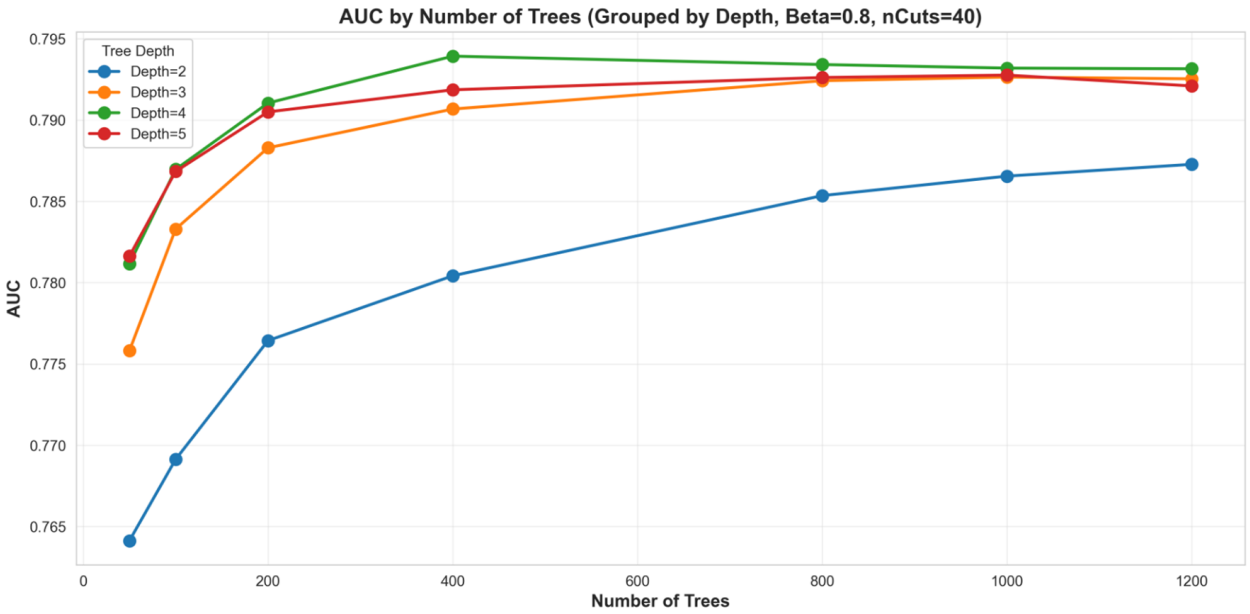

Рисунок 3.8 — Пример, иллюстрирующий анализ гиперпараметров (наблюдаемые из эксперимента ATLAS, 8 ГэВ)

Был проведен анализ гиперпараметров BDT с целью выбора оптимальных настроек модели. Была проведена сеточная оптимизация по следующим гиперпараметрам:

- Количество деревьев в ансамбле (NTrees) (50, 100, 200, 400, 800, 1000, 1200, 1400, 1600);
- Максимальная глубина каждого дерева (MaxDepth) (от 2 до 5 с шагом 1);
- Скорость обучения (Beta) (0.3, 0.5, 0.8);
- Количество разбиений по каждому признаку при построении дерева (nCuts) (20, 40);
- Тип бустинга (BoostType) (Grad, AdaBoost).

Выбор лучших гиперпараметров производился на основе метрики ROC AUC на тестовой выборке.

На основе проведённого анализа были выбраны оптимальные гиперпараметры для каждого энергетического диапазона. ROC-кривые и BDT-оценки классификатора для каждого энергетического диапазона приведены ниже на рисунках 3.9, 3.11, 3.13 (ROC-кривые) и 3.10, 3.12, 3.14 (BDT-оценки) соответственно.

Контроль переобучения для данных классификаторов выполнялся сравнением эффективности сигнала на тренировочной и тестовой выборках при

фиксированных уровнях режекции фона; отклонение метрик не превышает долей процента и не демонстрирует систематического превосходства тренировочной выборки, что свидетельствует об отсутствии значимого переобучения.

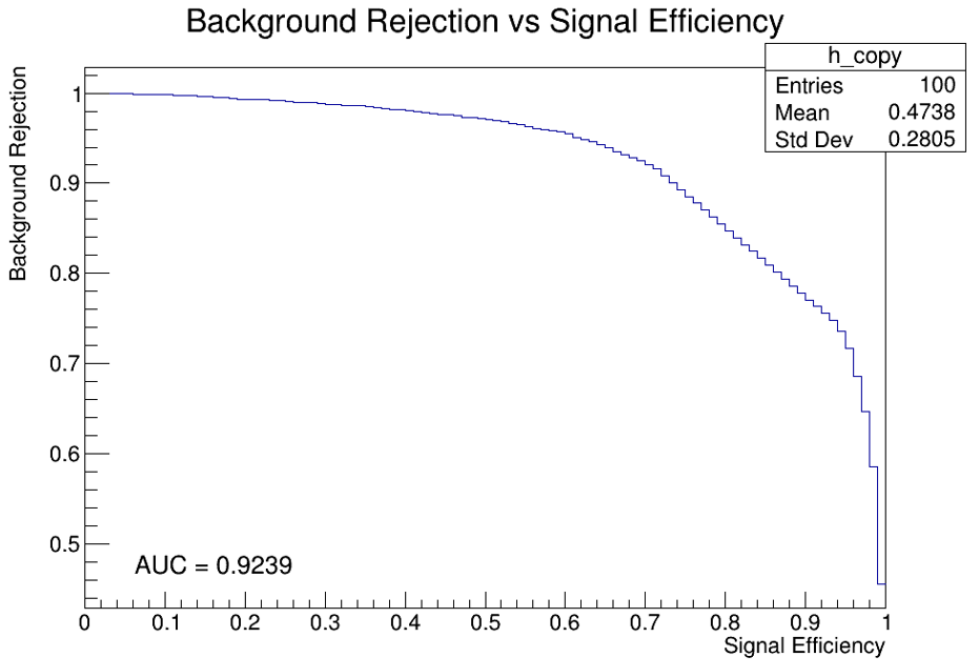

Рисунок 3.9 — ROC-кривая BDT-классификатора, полученная для случая: γ/π^0, запущенных с энергией 2 ГэВ; использованы наблюдаемые из эксперимента ATLAS; подобраны лучшие гиперпараметры

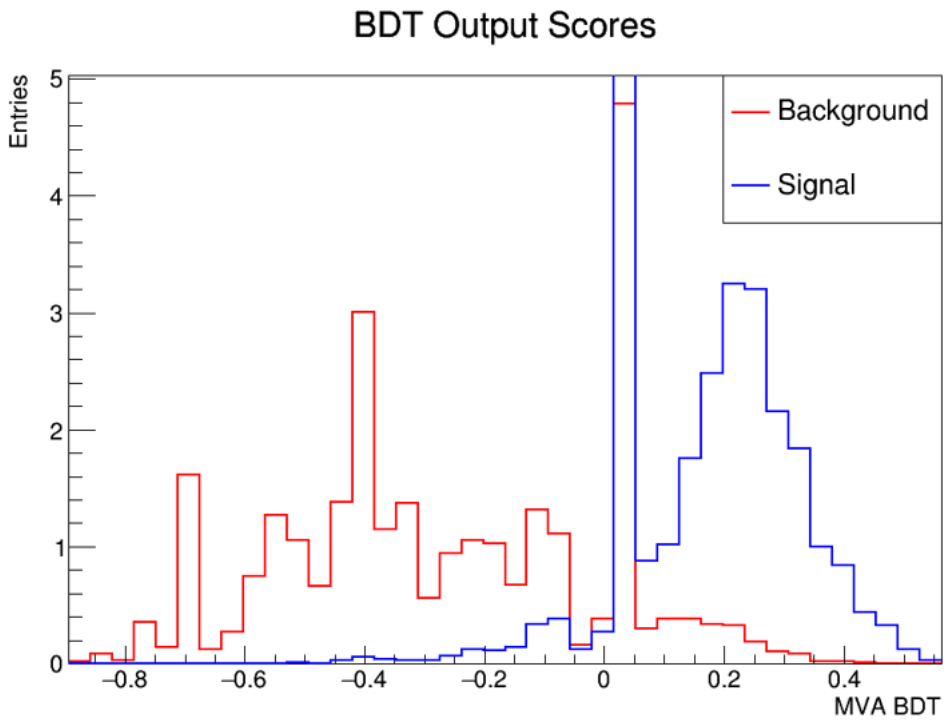

Рисунок 3.10 — BDT-оценки классификатора, полученные для случая: γ/π^0, запущенных с энергией 2 ГэВ; использованы наблюдаемые из эксперимента ATLAS; подобраны лучшие гиперпараметры

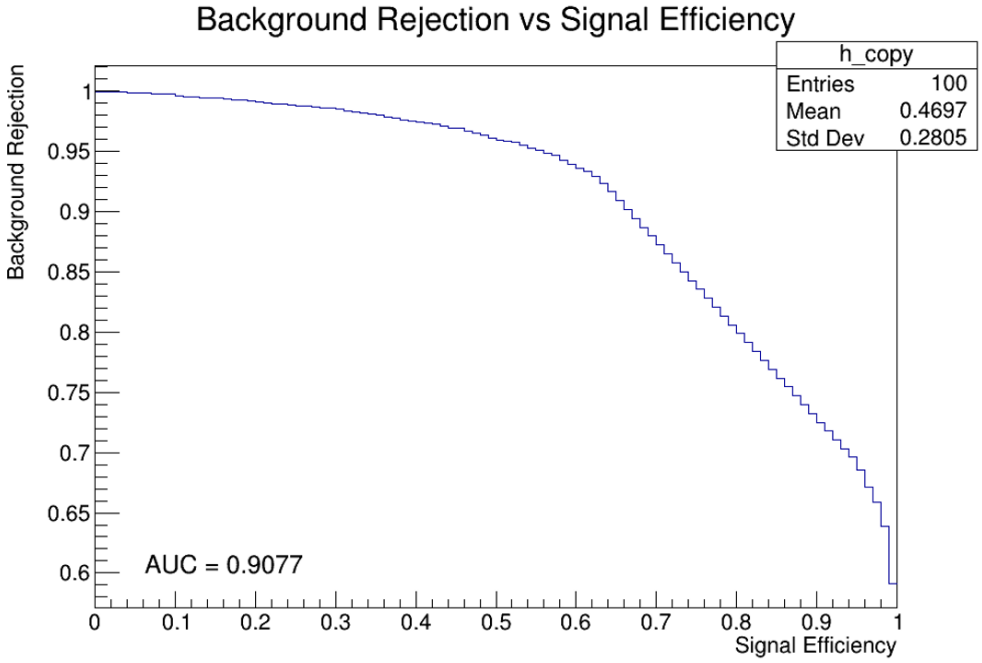

Рисунок 3.11 — ROC-кривая BDT-классификатора, полученная для случая: γ/π^0, запущенных с энергией 4 ГэВ; использованы наблюдаемые из эксперимента ATLAS; подобраны лучшие гиперпараметры

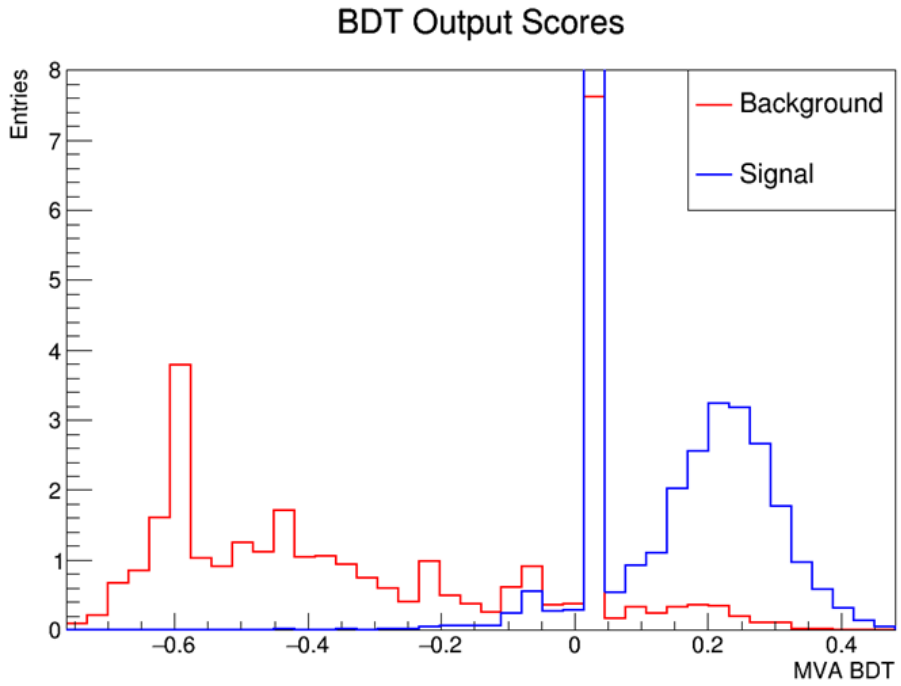

Рисунок 3.12 — BDT-оценки классификатора, полученные для случая: γ/π^0, запущенных с энергией 4 ГэВ; использованы наблюдаемые из эксперимента ATLAS; подобраны лучшие гиперпараметры

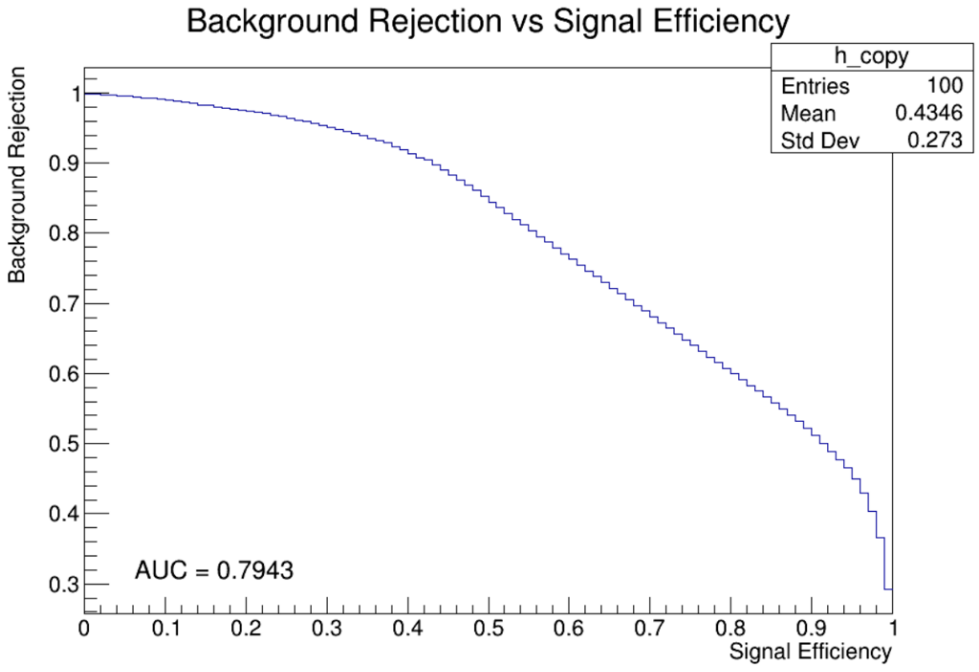

Рисунок 3.13 — ROC-кривая BDT-классификатора, полученная для случая: γ/π^0, запущенных с энергией 8 ГэВ; использованы наблюдаемые из эксперимента ATLAS; подобраны лучшие гиперпараметры

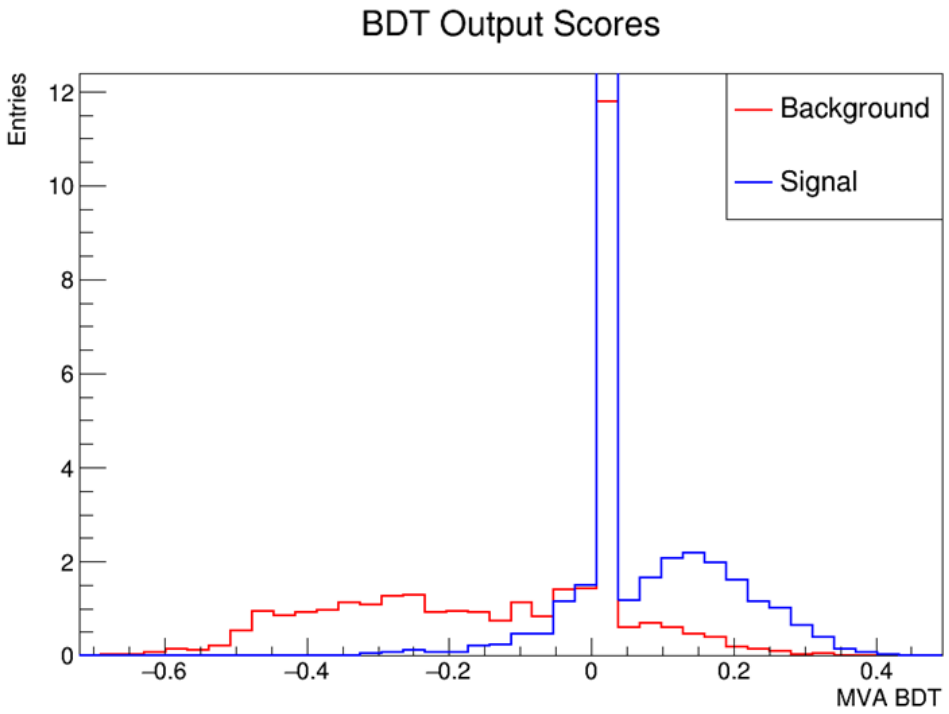

Рисунок 3.14 — BDT-оценки классификатора, полученные для случая: γ/π^0, запущенных с энергией 8 ГэВ; использованы наблюдаемые из эксперимента ATLAS; подобраны лучшие гиперпараметры

Из ROC-кривых, представленных на рисунках 3.9, 3.11 и 3.13, можно заключить, что BDT классификатор тем хуже разделяет фотонные и π^0-кластеры, чем выше энергия частиц. Данные результаты согласуются с ожиданиями и предыдущими исследованиями в данной области [6]. Это вероятно связано с тем, что при увеличении энергии расстояние между фотонами от распада π^0 уменьшается, и максимумы начинают сливаться, что затрудняет их разделение.

На рисунках 3.10, 3.12 и 3.14 можно заметить ярко выраженные пики. Их природа объяснена в следующей секции.

Важно также отметить, что полученные на данном этапе метрики (ROC AUC), являются сильно заниженными из-за отсутствия отбора на подтип кластера для π^0 (поскольку кластеры от π^0 подтипа [π^0,γ] имеют практически одинаковую форму с кластерами от γ) и других причин (см. следующую секцию).

3.2.2 КЛАССИФИКАЦИЯ МЕТОДОМ ФИКСИРОВАННЫХ ОТБОРОВ С ИСПОЛЬЗОВАНИЕМ АДАПТИРОВАННЫХ НАБЛЮДАЕМЫХ ДЛЯ КЛАСТЕРОВ ИЗ ЭКСПЕРИМЕНТА ATLAS

Для сравнения с BDT классификатором был реализован классификатор, основанный на фиксированных отборах, с использованием тех же адаптированных наблюдаемых для кластеров и той же выборки. С помощью TMVA были подобраны оптимальные пороговые значения для каждой наблюдаемой с целью максимизации режекции фона при различных эффективностях. В результате была построена ROC-кривая для классификатора, основанного на фиксированных отборах.

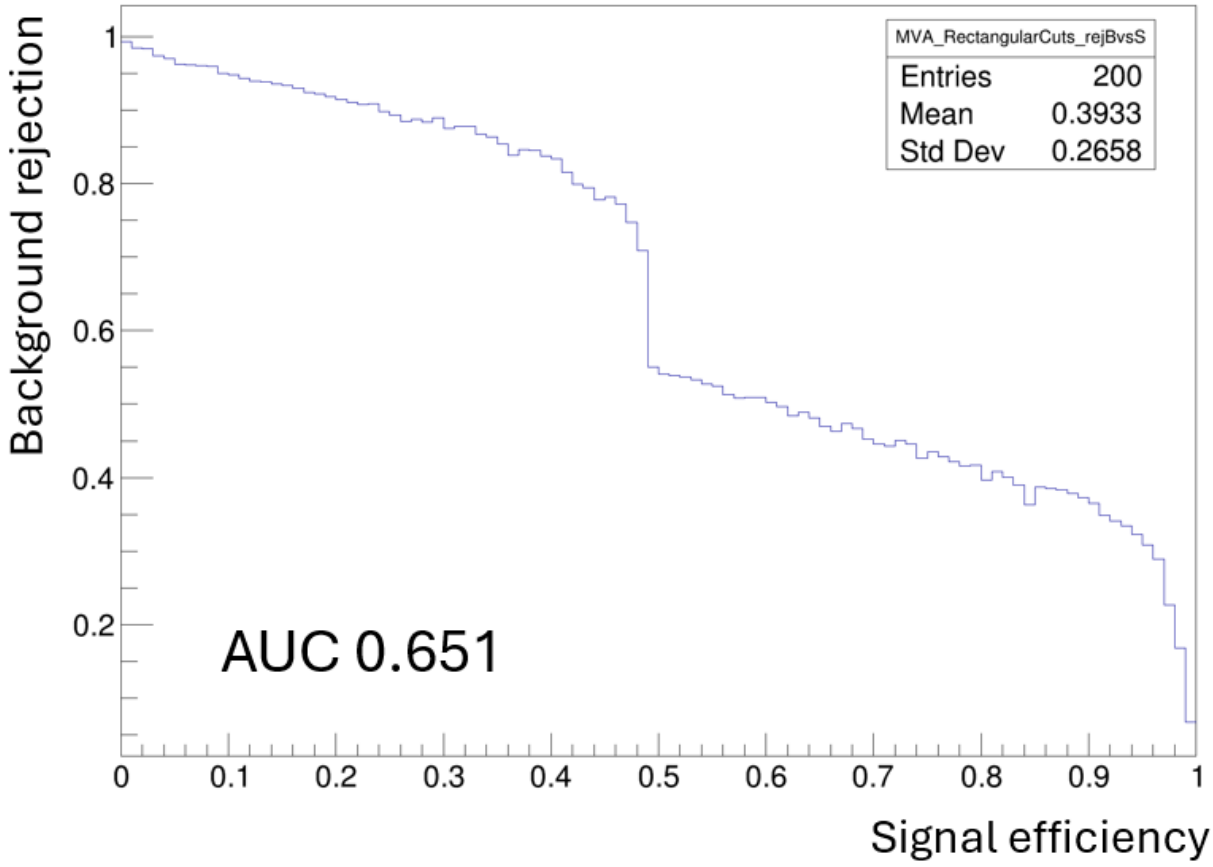

Рисунок 3.15 — ROC-кривая классификатора фиксированных отборов до удаления кластеров из одного хита, полученная для случая: γ/π^0, запущенных с энергией 8 ГэВ; использованы наблюдаемые из эксперимента ATLAS

Анализ ROC-кривых показал присутствие резкого порога в режекции фона при достижении определённой эффективности сигнала (рисунок 3.15). Была выдвинута гипотеза, что это связано с кластерами, состоящими из одного хита (ячейки калориметра). Таких кластеров порядка 20% от выборки. Данные кластеры имеют специфические значения наблюдаемых, что может приводить к резкому изменению эффективности при прохождении определённых порогов. Для проверки этой гипотезы была проведена дополнительная классификация с исключением кластеров из одного хита (рисунок 3.16).

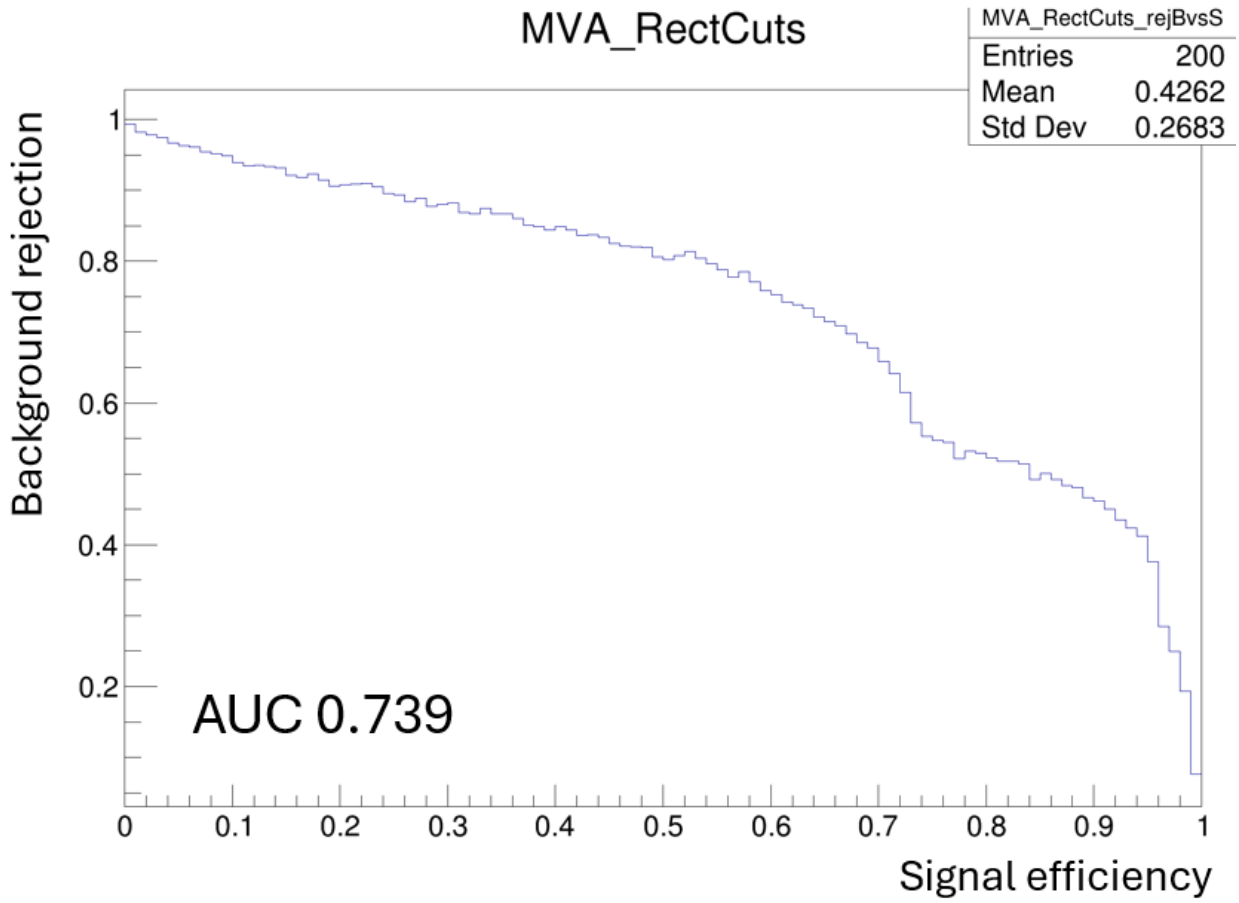

Рисунок 3.16 — ROC-кривая классификатора фиксированных отборов после удаления кластеров из одного хита, полученная для случая: γ/π^0, запущенных с энергией 8 ГэВ; использованы наблюдаемые из эксперимента ATLAS

При исключении таких кластеров резкий порог в ROC-кривой исчезает, что подтверждает гипотезу. Сложно представить себе модель, способную разделять такие кластеры, поскольку они не несут информации о форме распределения энергии. Поэтому всюду далее такие кластеры исключались из

анализа. После обучения классификаторов BDT на выборке с исключением кластеров из одного хита было также определено, что пики на BDT оценках на рисунках 3.10, 3.12 и 3.14 связаны именно с кластерами из одного хита. Кроме того было выявлено, что присутствие данных кластеров сильно занижает метрику ROC AUC как классификатора BDT, так и классификатора основанного на фиксированных отборах.

В сравнении с BDT классификатором классификатор, основанный на фиксированных отборах, показал результаты значительно хуже: так на 8 ГэВ при эффективности сигнала 80% режекция фона составила всего около 40%, в то время как у BDT он был около 60% (сравнение до исключения кластеров из одного хита).

3.2.3 ИСПОЛЬЗОВАНИЕ ВСЕЙ СОВОКУПНОСТИ НАБЛЮДАЕМЫХ ИЗ ATLAS И ТЕКУЩЕГО АЛГОРИТМА КЛАССИФИКАЦИИ В SPDROOT ДЛЯ КЛАССИФИКАЦИИ КЛАСТЕРОВ С ПОМОЩЬЮ BDT

Для дальнейшего улучшения классификации кластеров был проведён анализ с использованием всей совокупности наблюдаемых из ATLAS и текущего алгоритма классификации в SPDROOT (MLP). Были использованы все наблюдаемые, представленные в SPDROOT. При этом были внесены небольшие незначительные изменения, упрощающие их вычисление с использованием уже реализованных функций. Подробное описание адаптации наблюдаемых ((x, y) соответствуют (η, φ)):

- **Смещение центра тяжести относительно максимума η/φ:**

$$x_{\text{cog}} = \frac{1}{S_{25}} \sum_{i=1}^{25} E_i \, X_i^{\text{rel}}, \qquad y_{\text{cog}} = \frac{1}{S_{25}} \sum_{i=1}^{25} E_i \, Y_i^{\text{rel}},$$

где $S_{25} = \sum_{i=1}^{25} E_i$, X_i^{rel}, Y_i^{rel} — координаты i-го хита относительно максимума.

- **Вторые моменты:**

$$\langle r^2 \rangle = S_{XX} + S_{YY} = \frac{\sum_{i=1}^{N} e_i \left[(x_i - x_c)^2 + (y_i - y_c)^2 \right]}{\sum_{i=1}^{N} e_i}.$$

где N=25, e_i — энергия i-го хита, (x_c, y_c) — координаты центра тяжести кластера ($\eta_c = \frac{\sum_{i=1}^{N} E_i \eta_i}{\sum_{i=1}^{N} E_i}$, $\phi_c = \arctan\left(\frac{\sum_{i=1}^{N} E_i \sin \phi_i}{\sum_{i=1}^{N} E_i \cos \phi_i} \right)$).

$$S_{XX} = \frac{\sum_{i=1}^{N} e_i (x_i - x_c)^2}{\sum_{i=1}^{N} e_i}, \qquad S_{YY} = \frac{\sum_{i=1}^{N} e_i (y_i - y_c)^2}{\sum_{i=1}^{N} e_i},$$

$$S_{XY} = S_{YX} = \frac{\sum_{i=1}^{N} e_i (x_i - x_c)(y_i - y_c)}{\sum_{i=1}^{N} e_i}.$$

где N=25, e_i — энергия i-го хита, (x_c, y_c) — координаты центра тяжести кластера.

- **Наблюдаемая формы кластера:**

$$\kappa = \sqrt{1 - \frac{S_{XX} S_{YY} - S_{XY}^2}{(S_{XX} + S_{YY})^2}} = \sqrt{1 - \frac{4 \det S}{\mathrm{Tr}(S)^2}},$$

где $\det S = S_{XX} S_{YY} - S_{XY}^2$, $\mathrm{Tr}(S) = S_{XX} + S_{YY}$.

- **Энергетические отношения:**

$$\frac{S_1}{S_9}, \qquad \frac{S_9 - S_1}{S_{25} - S_1}, \qquad \frac{M_2 + S_1}{S_4}, \qquad \frac{S_6}{S_9}, \qquad \frac{M_2 + S_1}{S_9}.$$

где S_1 — максимальная энергия в ячейке кластера, M_2 — вторая по величине энергия в ячейке кластера, S_4, S_6, S_9, S_{25} — суммы энергий в окнах 2×2, 3×2, 3×3, 5×5 ячеек соответственно по (η, ϕ).

- **Важность хвостов (Tail Importance):**

$$\mathrm{TI} = \frac{\langle r^4 \rangle}{\langle r^2 \rangle^2} = 1 - \frac{\langle r^2 \rangle^2}{\langle r^4 \rangle}.$$

где

$$\langle r^4 \rangle = \frac{\sum_{i=1}^{N} e_i \left[(x_i - x_c)^2 + (y_i - y_c)^2 \right]^2}{\sum_{i=1}^{N} e_i}.$$

Были построены корреляционные матрицы для полного набора наблюдаемых для кластеров (рисунки 3.18, 3.17). Анализ показал наличие значительных корреляций между некоторыми наблюдаемыми, что позволяет высказать предположение, что возможно произвести отбор наблюдаемых без потери качества (или даже с ростом качества) классификации. В дальнейшем будет показано влияние данных наблюдаемых на эффективность классификации.

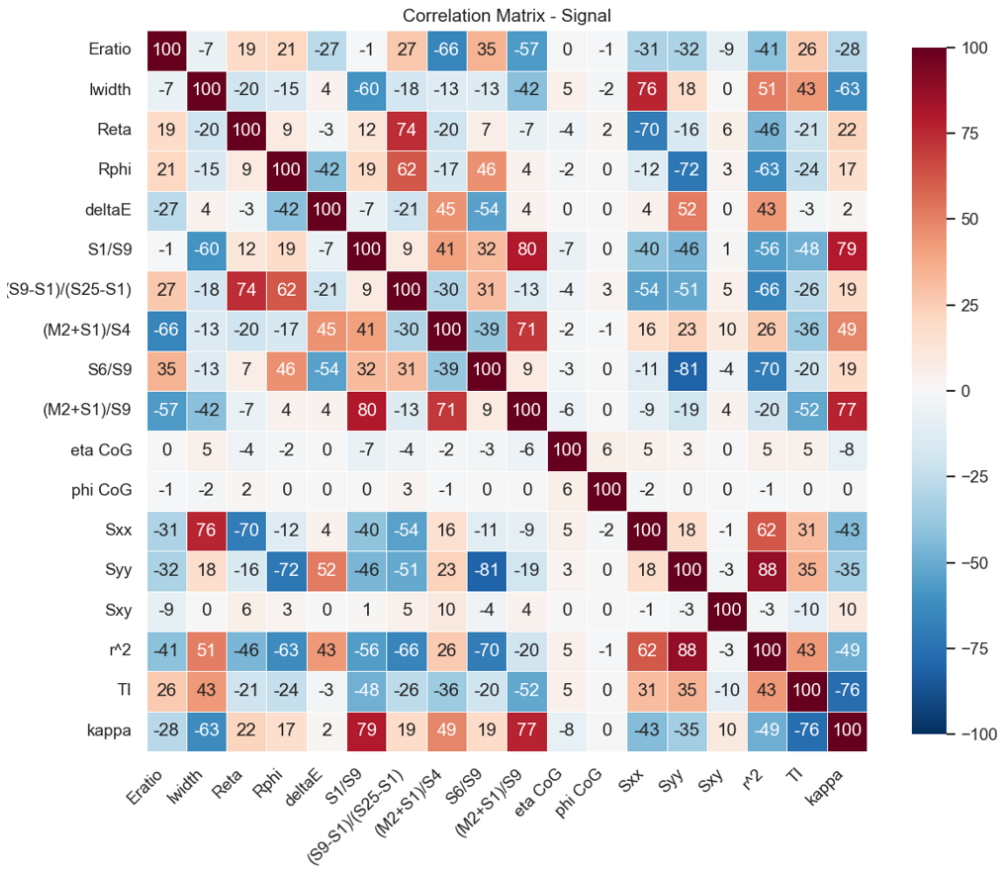

Рисунок 3.17 — Корреляционная матрица полного набора наблюдаемых для кластеров для сигналов (фотоны)

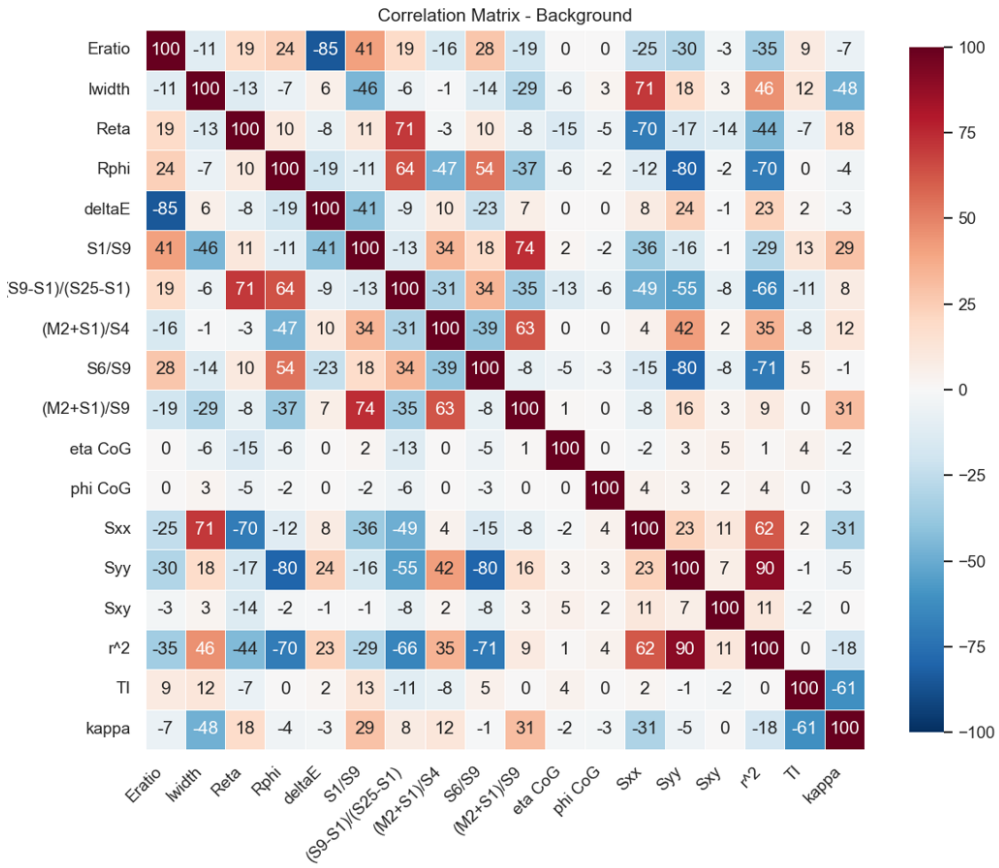

Рисунок 3.18 — Корреляционная матрица полного набора наблюдаемых для кластеров для фона (π^0)

Далее был обучен BDT классификатор с использованием полного набора наблюдаемых для кластеров на той же выборке для 2, 4 и 8 ГэВ. Были проведены оптимизации гиперпараметров аналогично предыдущему разделу. ROC-кривые и BDT-оценки классификатора для каждого энергетического диапазона приведены ниже на рисунках 3.19, 3.21, 3.23 (ROC-кривые) и 3.20, 3.22, 3.24 (BDT-оценки) соответственно.

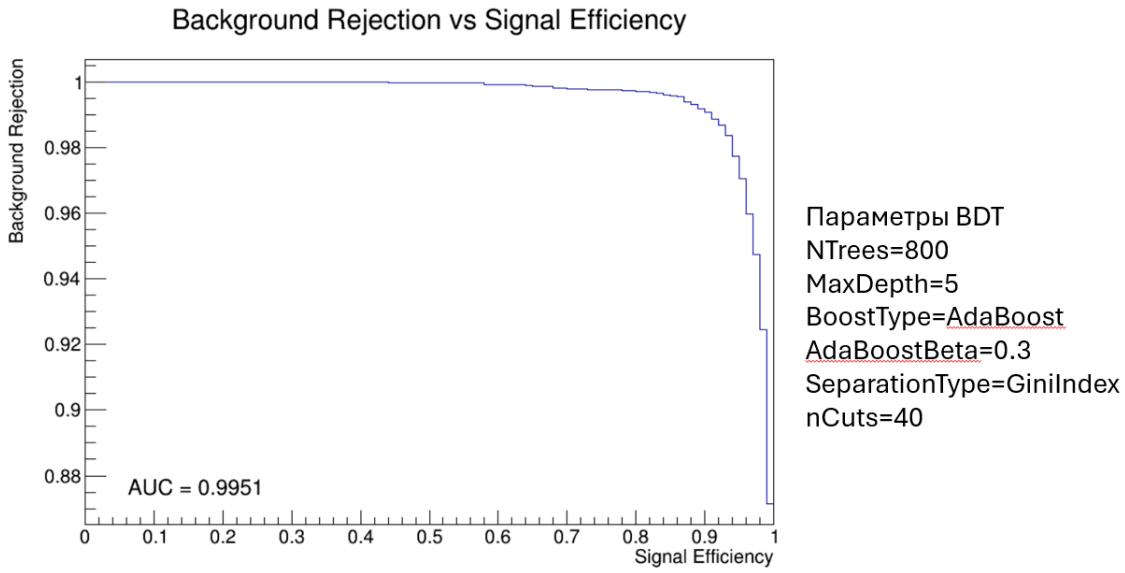

Рисунок 3.19 — ROC-кривая BDT-классификатора, полученная для случая: γ/π^0, запущенных с энергией 2 ГэВ; использованы все наблюдаемые; подобраны лучшие гиперпараметры

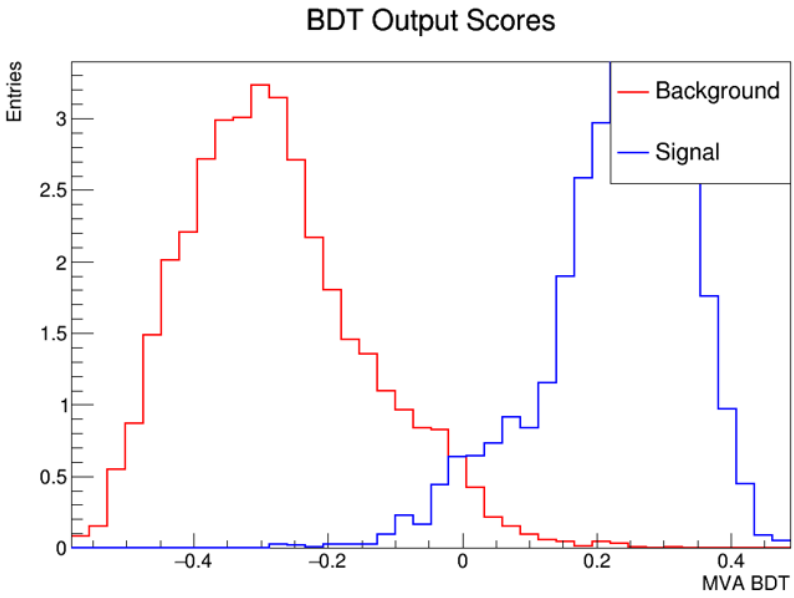

Рисунок 3.20 — BDT-оценки классификатора, полученные для случая: γ/π^0, запущенных с энергией 2 ГэВ; использованы все наблюдаемые; подобраны лучшие гиперпараметры

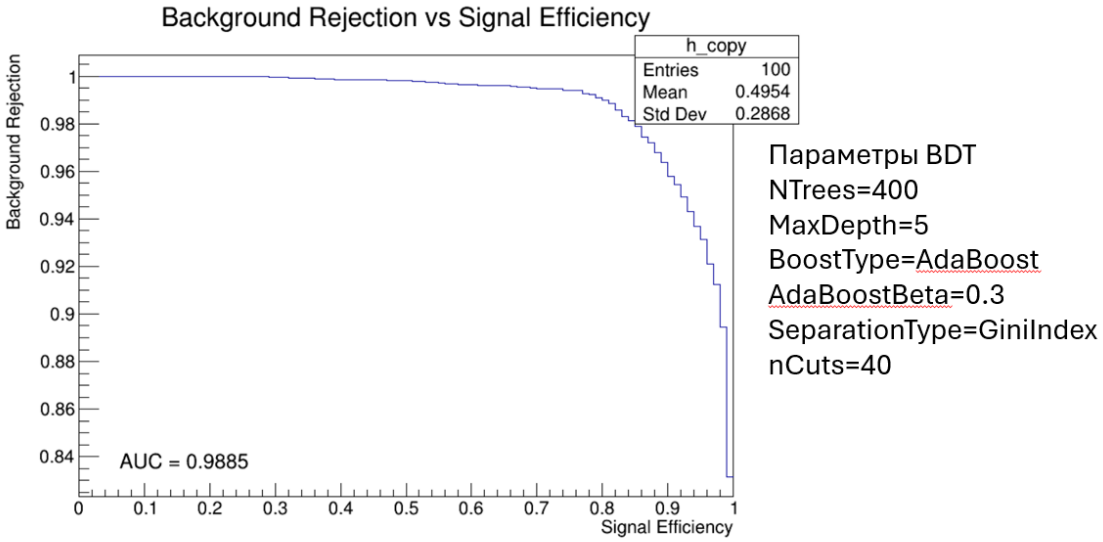

Рисунок 3.21 — ROC-кривая BDT-классификатора, полученная для случая: γ/π^0, запущенных с энергией 4 ГэВ; использованы все наблюдаемые; подобраны лучшие гиперпараметры

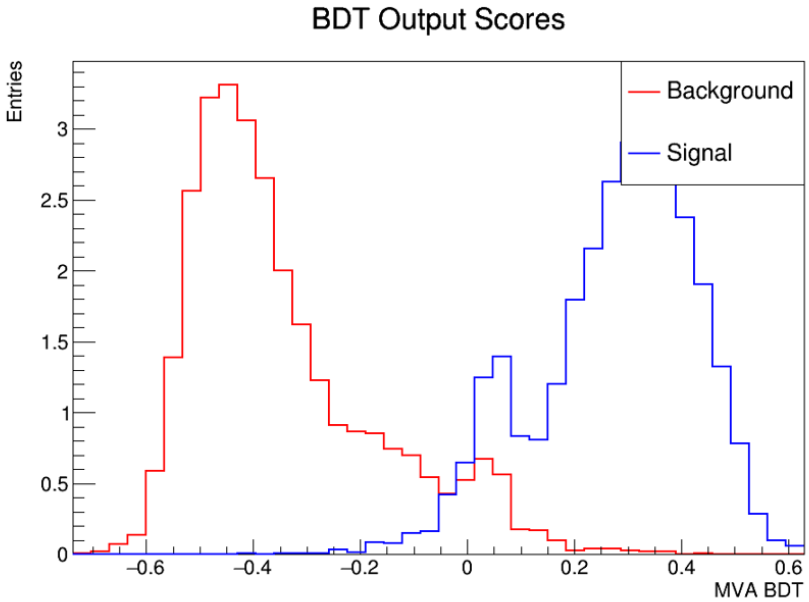

Рисунок 3.22 — BDT-оценки классификатора, полученные для случая: γ/π^0, запущенных с энергией 4 ГэВ; использованы все наблюдаемые; подобраны лучшие гиперпараметры

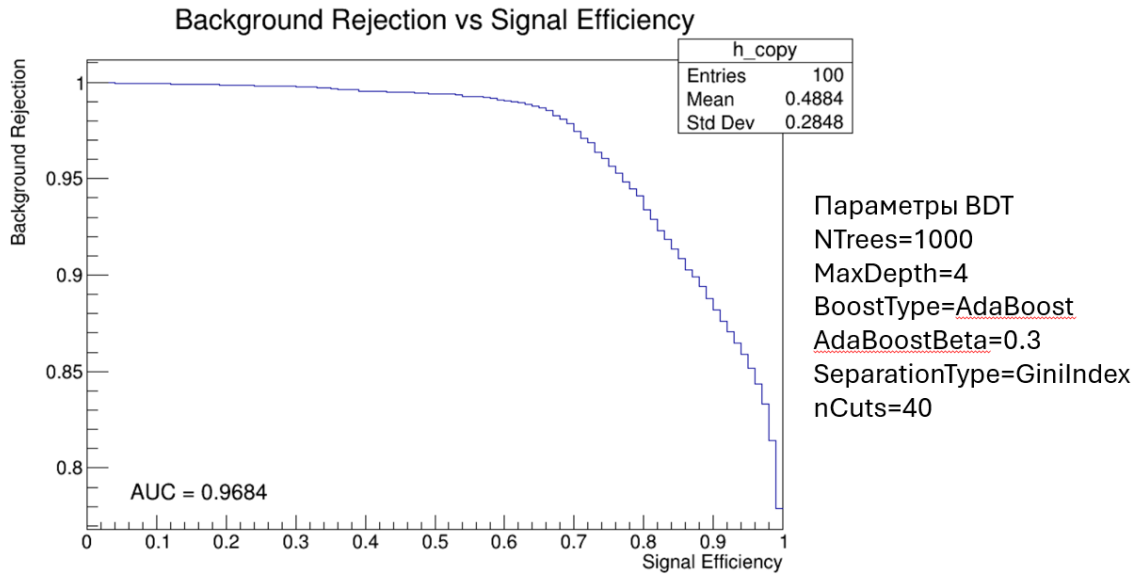

Рисунок 3.23 — ROC-кривая BDT-классификатора, полученная для случая: γ/π^0, запущенных с энергией 8 ГэВ; использованы все наблюдаемые; подобраны лучшие гиперпараметры

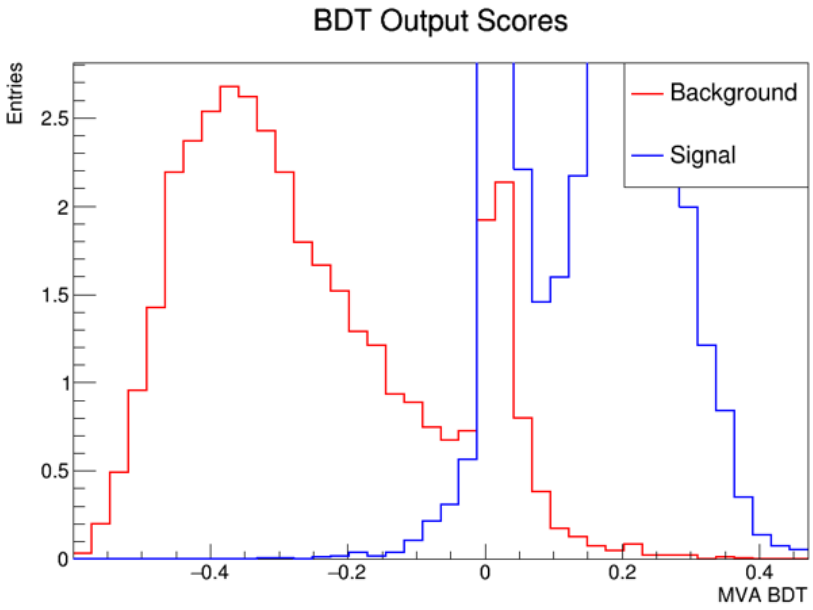

Рисунок 3.24 — BDT-оценки классификатора, полученные для случая: γ/π^0, запущенных с энергией 8 ГэВ; использованы все наблюдаемые; подобраны лучшие гиперпараметры

Из анализа ROC-кривых (рисунки 3.19, 3.21, 3.23) можно заключить, что данный BDT классификатор показывает улучшенные результаты по сравнению с предыдущими версиями, особенно на высоких энергиях (4 и 8 ГэВ). Для эффективности сигнала около 80% режекция фона достигает примерно

95% на 8 ГэВ и 99.5% на 2 ГэВ, что значительно лучше предыдущих результатов.

Улучшение связано с исключением кластеров из одного хита и добавлением дополнительных наблюдаемых. Тем не менее метрики здесь полученные, по-прежнему занижены в связи с отсутствием отбора на подтип кластера.

3.3 ТЕСТИРОВАНИЕ КЛАССИФИКАТОРОВ НА БОЛЕЕ РЕАЛИСТИЧНЫХ ВЫБОРКАХ

3.3.1 УТОЧНЕНИЕ ПОСТАНОВКИ ЗАДАЧИ

В данной секции по ранее описанным причинам для улучшения качества классификации на реалистичных выборках необходимо уточнить постановку задачи касательно подтипов кластеров.

Для кластеров от π^0 существует 2 принципиальных случая (другие типы кластеров маловероятны либо будут идентифицироваться другими детекторами):

1) Фотоны от π^0 попали в один кластер (условное обозначение $[\pi^0, \gamma, \gamma]$): тогда такой кластер имеет специфическую форму
2) Фотоны от π^0 попали в разные кластеры (условное обозначение $[\pi^0, \gamma]$): такие кластеры можно идентифицировать из кинематики

В дальнейшем в качестве фона рассмотрены кластеры 1го типа, именно такие кластеры отделялись от «фотонных кандидатов»: кластеров 2го типа и кластеров от прямых фотонов (для прямых фотонов брались только кластеры, в которых был один фотон). Соответственно рассмотрены будут 2 сценария разделения:

1) фон: кластеры 1го типа; сигнал: кластеры от прямых фотонов
2) фон: кластеры 1го типа; сигнал: кластеры 2го типа и кластеры от прямых фотонов

3.3.2 КАЧЕСТВО КЛАССИФИКАТОРОВ ПО МЕРЕ ПРИБЛИЖЕНИЯ ВЫБОРКИ К БОЛЕЕ РЕАЛИСТИЧНОМУ СЛУЧАЮ

Тестирование проводилось в 3 этапа (на всех этапах разбиение выборки на тренировочную и тестовую производилось случайным образом 70% на 30%):

1) запуск одиночных γ/π^0 перпендикулярно цилиндрической части калориметра с равномерным распределением по поперечному импульсу от 0 до 10 ГэВ и отбором на энергию кластера >1 ГэВ; выборка (тренировочная) сигнала: 13867 кластеров; выборка (тренировочная) фона: 7457 кластеров; рассматривалась только цилиндрическая часть калориметра (далее этап 1)
2) запуск одиночных γ/π^0 с равномерным распределением по телесному углу и равномерным распределением по поперечному импульсу от 0 до 10 ГэВ и отбором на энергию кластера >1 ГэВ; выборка (тренировочная) сигнала: 12765 кластеров; выборка (тренировочная) фона: 8293 кластеров; рассматривалась только цилиндрическая часть калориметра (далее этап 2)
3) minbias (события без отбора по конкретному физическому процессу, сгенерированные с помощью PYTHIA8 [11]) pp (протон-протонные) столкновения с энергией в системе центра масс $\sqrt{s} = 27$ ГэВ и с отбором на энергию кластера >1 ГэВ; выборка (тренировочная) сигнала: 29746 кластеров; выборка (тренировочная) фона: 119319 кластеров; моделирование включало как цилиндрическую, так и торцевую части (эндкапы) калориметра (далее этап 3)

Для BDT классификаторов использовались 18 наблюдаемых, описанных ранее (наблюдаемые из эксперимента ATLAS + наблюдаемые из SPDROOT) и 4 наблюдаемые, отвечающие энергии и положению кластера в пространстве: $E_{\text{cluster}} = \sum_i E_i$ — полная энергия кластера; η_{central} — псевдобыстрота центрального хита (хита, ближайшего к энергетическому центру кластера); ϕ_{central} — азимутальный угол центрального хита; z_{central} — z-координата центрального хита. Для каждого этапа была проведена оптимизация гиперпараметров (аналогично оптимизации в предыдущих секциях). Результаты тестирования классификаторов на более реалистичных выборках для сценария 1

приведены на рисунках 3.25, 3.26 и 3.27.

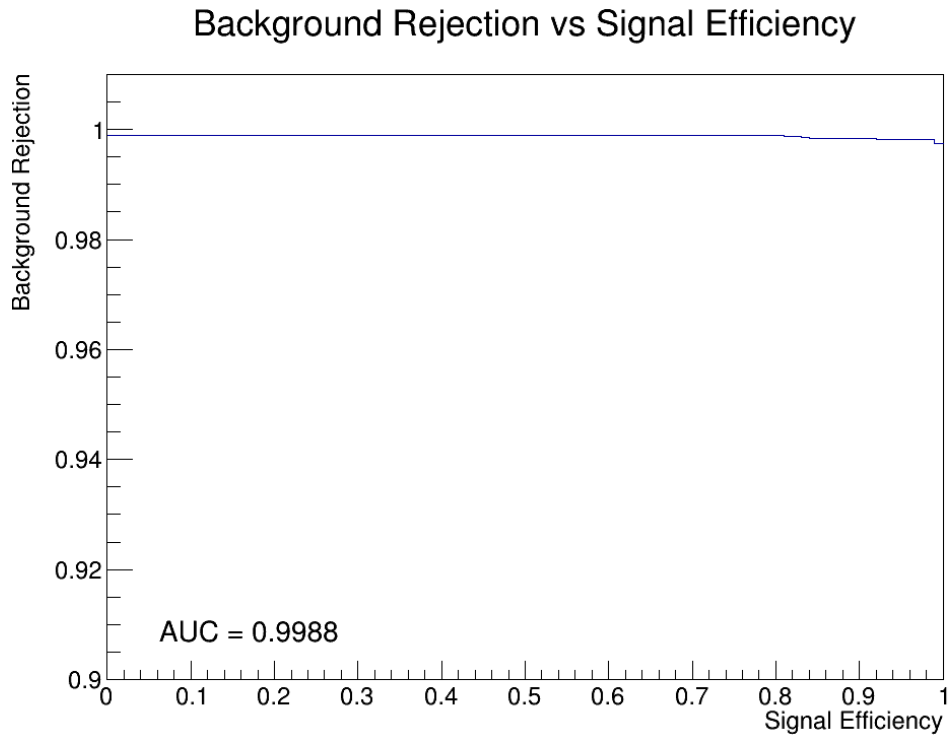

Рисунок 3.25 — ROC-кривая BDT классификатора (использованы лучшие гиперпараметры: BoostType = grad, nTrees = 800, MaxDepth = 5, nCuts = 40) для этапа 1

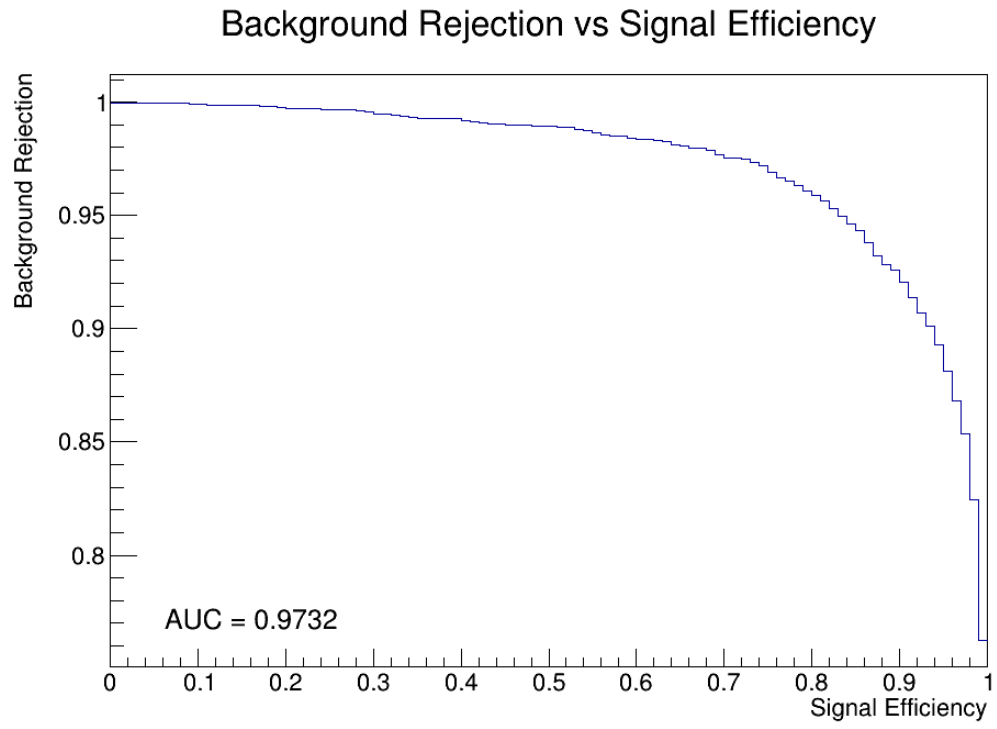

Рисунок 3.26 — ROC-кривая BDT классификатора (использованы лучшие гиперпараметры: BoostType = AdaBoost, nTrees = 1400, MaxDepth = 5, nCuts = 40) для этапа 2

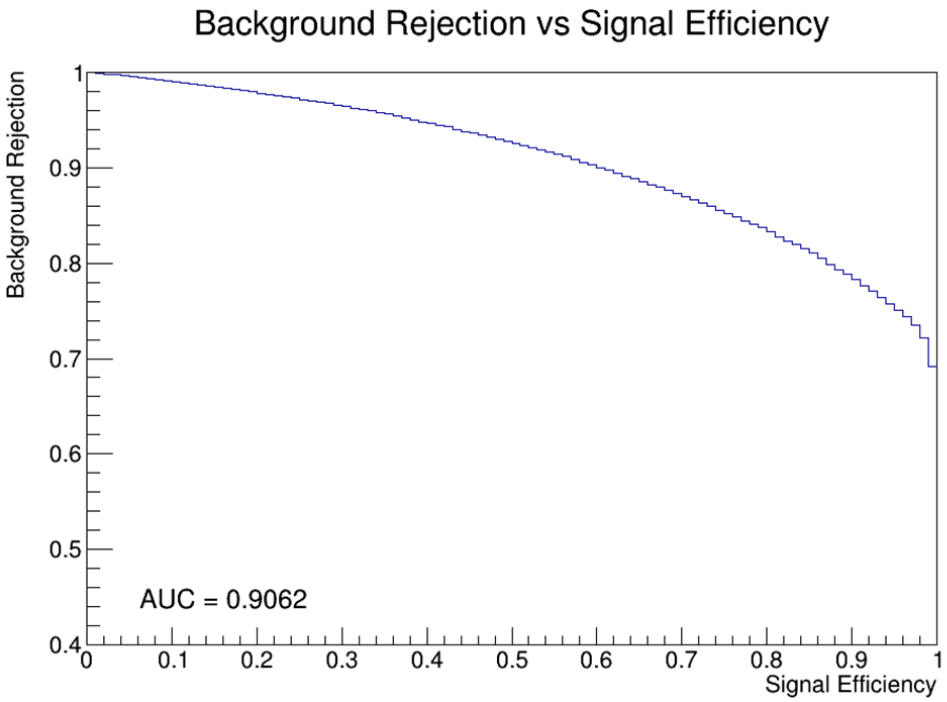

Рисунок 3.27 — ROC-кривая BDT классификатора (использованы лучшие гиперпараметры: BoostType = Grad, nTrees = 200, MaxDepth = 2, nCuts = 40) для этапа 3

Как видно из представленных ROC-кривых (рис. 3.25–3.27), по мере усложнения выборки от этапа 1 к этапу 3 качество классификации закономерно снижается. ROC AUC падает с 0.999 (этап 1) до 0.973 (этап 2) и 0.906 (этап 3). Падение режекции фона обусловлено усложнением топологии событий: появлением перекрывающихся ливней и разбросом углов падения частиц. Даже на выборке minbias pp-столкновений BDT-классификатор сохраняет хорошую разделяющую способность (ROC AUC = 0.906), что свидетельствует о перспективности данного подхода для задач идентификации фотонов в эксперименте SPD.

3.3.3 СРАВНЕНИЕ СЦЕНАРИЕВ РАЗДЕЛЕНИЯ

Было проведено сравнение метрик BDT-классификаторов для описанных выше сценариев разделения. Для первого сценария (сигнал: кластеры от прямых фотонов; фон: кластеры 1-го типа) ROC-кривая была представлена на рисунке 3.26. Для второго сценария (сигнал: фотонные кандидаты от π^0 и кластеры от прямых фотонов; фон: кластеры 1-го типа) была получена ROC-кривая (рисунок 3.28):

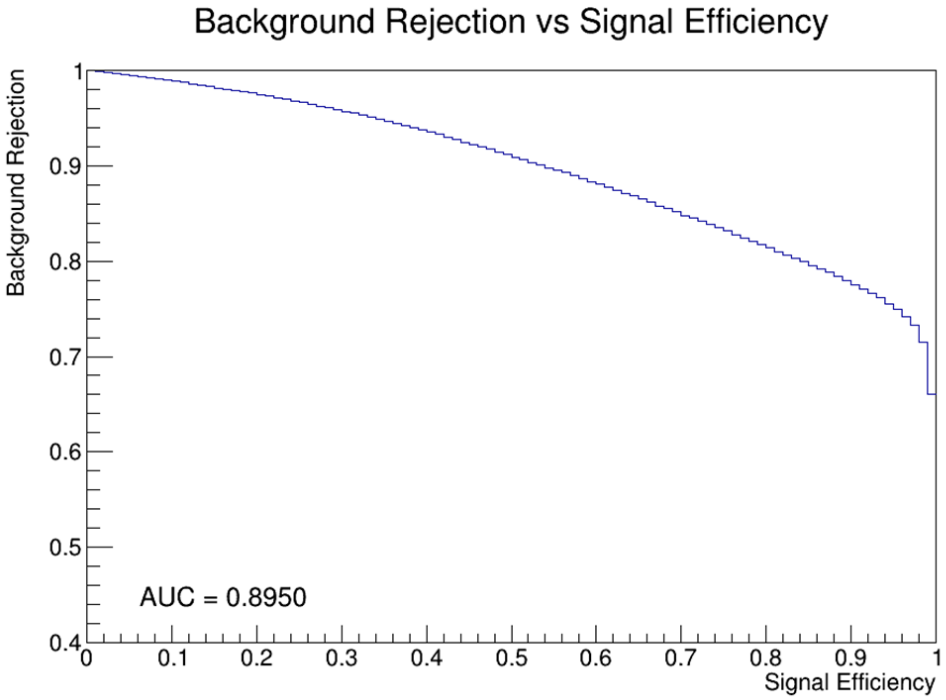

Рисунок 3.28 — ROC-кривая BDT классификатора для второго сценария разделения для этапа 3

Как видно из ROC-кривых, добавление фотонных кандидатов практически не влияет на эффективность классификатора (влияние порядка 1%). Это ожидаемо, поскольку кластеры от различных типов сигнала имеют схожую форму.

3.4 ОТБОР ЛУЧШИХ НАБЛЮДАЕМЫХ

3.4.1 РАНЖИРОВАНИЕ НАБЛЮДАЕМЫХ ДЛЯ ОПТИМИЗАЦИИ КЛАССИФИКАТОРА BDT

Для ранжирования наблюдаемых использовался метод N-1 ранжирования. Был реализован итеративный алгоритм исключения наблюдаемых:

1) Обучается референтная (базовая) модель BDT на всех N (22) наблюдаемых, вычисляется AUC_{all} (ROC AUC референтной модели).
2) Для каждой наблюдаемой x_i из текущего набора:
   (а) Обучается модель без x_i и всех ранее исключённых наблюдаемых;
   (б) Вычисляется AUC_{without} (ROC AUC модели из п. а);
   (в) Рассчитывается накопленное влияние (impact):

$$\text{Impact}^{(k)}(x_i) = \frac{AUC_{\text{all}} - AUC_{\text{all}\setminus(R_{k-1}\cup\{x_i\})}}{AUC_{\text{all}}} \times 100\%,$$

где k — номер итерации; R_{k-1} — множество наблюдаемых, исключённых на предыдущих $k-1$ итерациях; AUC_{all} — метрика модели со всеми 22 наблюдаемыми; $AUC_{\text{all}\setminus(R_{k-1}\cup\{x_i\})}$ — метрика модели без x_i и всех ранее исключённых наблюдаемых.

3) Наблюдаемая с наименьшим impact исключается.
4) Шаги 2–3 повторяются, пока не останется одна наблюдаемая.

Таким образом, на k-й итерации impact отражает накопленный вклад k исключённых наблюдаемых (всех предыдущих плюс кандидата). Итеративное исключение проводилось на данных для этапа 3 (minbias) с уменьшенной в 5 раз выборкой. Ниже представлены накопленное влияние каждой наблюдаемой на момент исключения, а также значение ROC AUC по итерациям (рисунки 3.29 и 3.30).

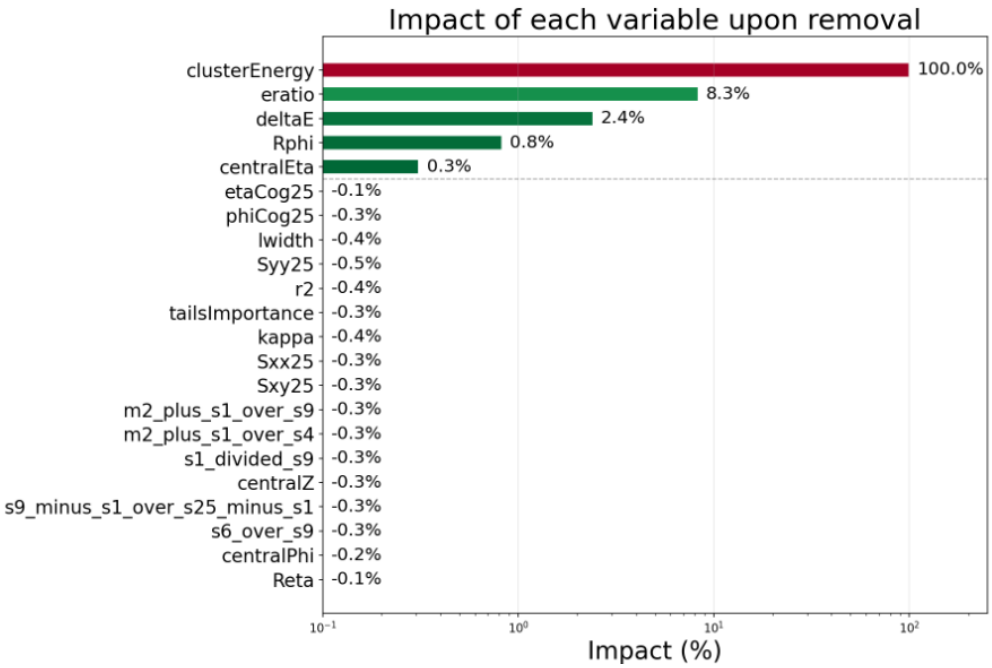

Рисунок 3.29 — Влияние исключения каждой наблюдаемой относительно референтной модели со всеми 22 наблюдаемыми

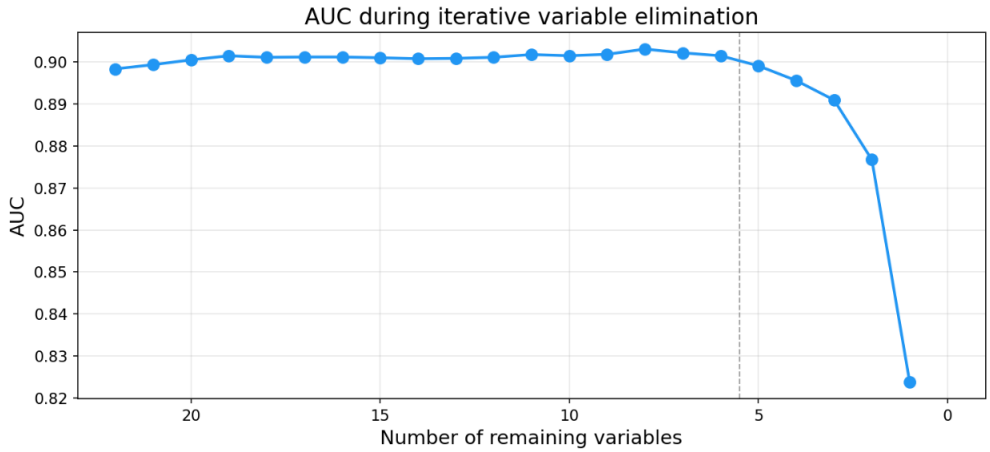

Рисунок 3.30 — ROC AUC в зависимости от числа оставшихся наблюдаемых

На основе анализа результатов итеративного исключения было определено, что для BDT классификатора можно сузить список наблюдаемых до 5 без значительной потери качества, поскольку далее ROC AUC выходит на плато. Отобранные наблюдаемые: E_{cluster}, E_{ratio}, ΔE, R_{ϕ}, η_{central}. ROC-кривая BDT классификатора после отбора 5 лучших наблюдаемых представлена ниже (рисунок 3.31).

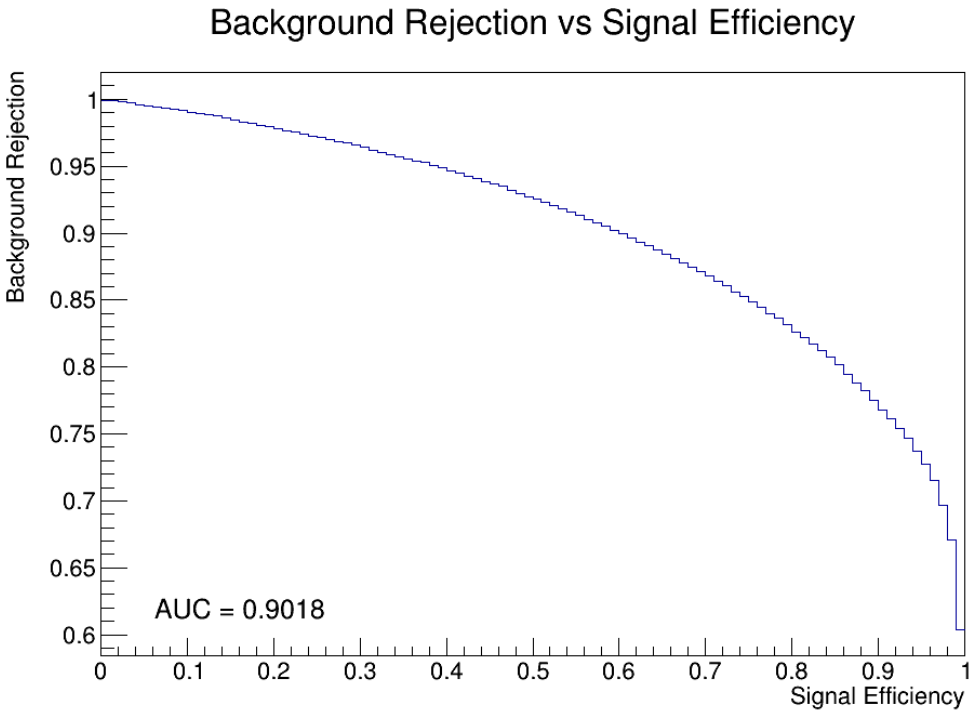

Рисунок 3.31 — ROC кривая BDT классификатора (с параметрами BoostType = Grad, nTrees = 200, MaxDepth = 2, nCuts = 40) для этапа 3, после отбора 5 лучших наблюдаемых

3.4.2 ОТБОР ЛУЧШИХ НАБЛЮДАЕМЫХ ДЛЯ КЛАССИФИКАТОРА ПРЯМОУГОЛЬНЫХ ФИКСИРОВАННЫХ ОТБОРОВ

Отбор лучших же наблюдаемых для классификатора прямоугольных отборов производился другим способом. В связи с крайне затратным по времени обучением классификатора на большом количестве наблюдаемых (и вообще говоря слабой устойчивости данного метода к добавлению новых наблюдаемых) вместо итеративного исключения использовался алгоритм итеративного добавления. Алгоритм последовательно добавлял наблюдаемые к изначально пустому набору:

1) На первом шаге обучалась модель с каждой наблюдаемой по отдельности; наблюдаемая, показавшая наибольшее ROC AUC, отбиралась первой.
2) На каждом последующем шаге для каждой из оставшихся наблюдаемых обучалась модель с добавлением данной наблюдаемой к текущему набору. Рассчитывался прирост качества (Gain):

$$\text{Gain} = \frac{AUC_{\text{new}} - AUC_{\text{prev}}}{AUC_{\text{prev}}} \times 100\%,$$

   где AUC_{prev} — метрика на предыдущем шаге, AUC_{new} — после добавления кандидата.
3) Если максимальный Gain среди всех кандидатов превышал порог 0.1%, соответствующая наблюдаемая добавлялась в набор; в противном случае алгоритм останавливался.
4) Шаги 2–3 повторялись, пока не оставалось кандидатов с Gain выше порога.

Отбор проводился среди 18 наблюдаемых формы кластера (ATLAS + SPDROOT) на данных этапа 3 с уменьшенной в 5 раз выборкой.

Ниже же представлены ROC AUC по итерациям и Gain при добавлении различных наблюдаемых (рисунки 3.33 и 3.32 соответственно).

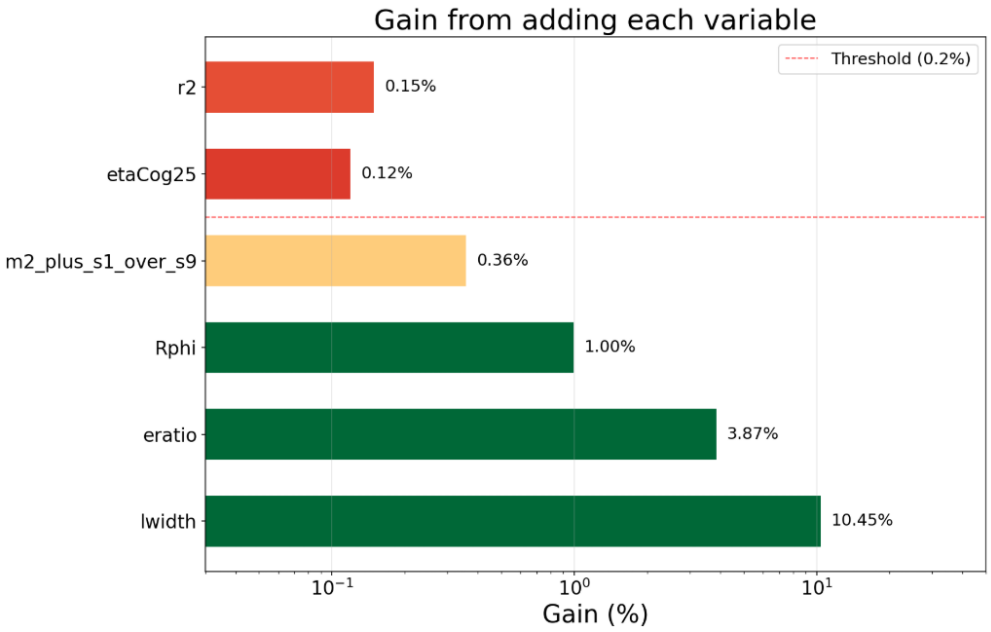

Рисунок 3.32 — Gain при добавлении наблюдаемых (кроме 1-й — ΔE) для классификатора прямоугольных отборов

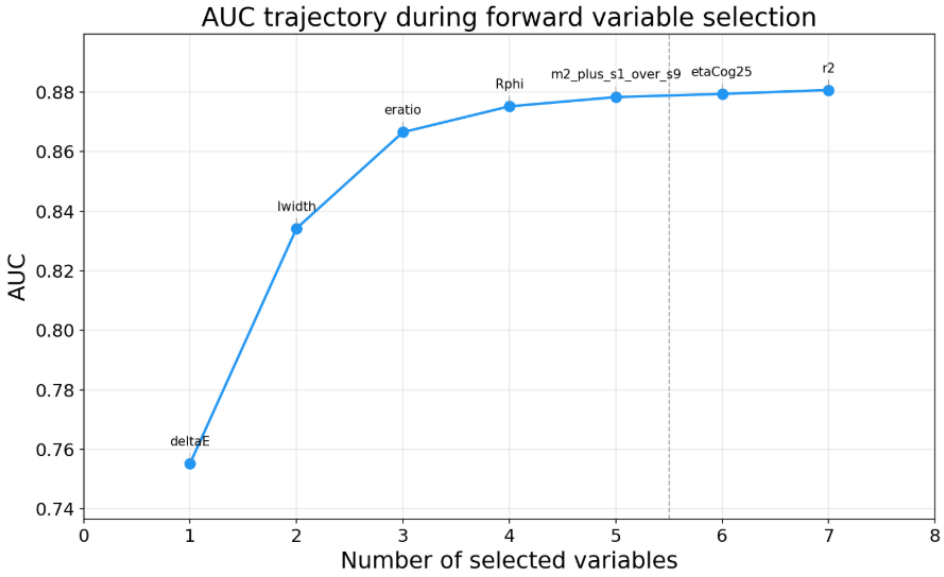

Рисунок 3.33 — ROC AUC в зависимости от числа добавленных наблюдаемых для классификатора прямоугольных отборов

Данный алгоритм остановился после добавления 7-й наблюдаемой. По порогу в 0.2% (далее ROC AUC выходит на плато) были выбраны 5 лучших наблюдаемых: E_{ratio}, $w_{\eta 2}$, R_{ϕ}, ΔE, $(M_2 + S_1)/S_9$. ROC-кривая классификатора прямоугольных отборов с данными наблюдаемыми для этапа 3 приведена на рисунке 3.34 ниже:

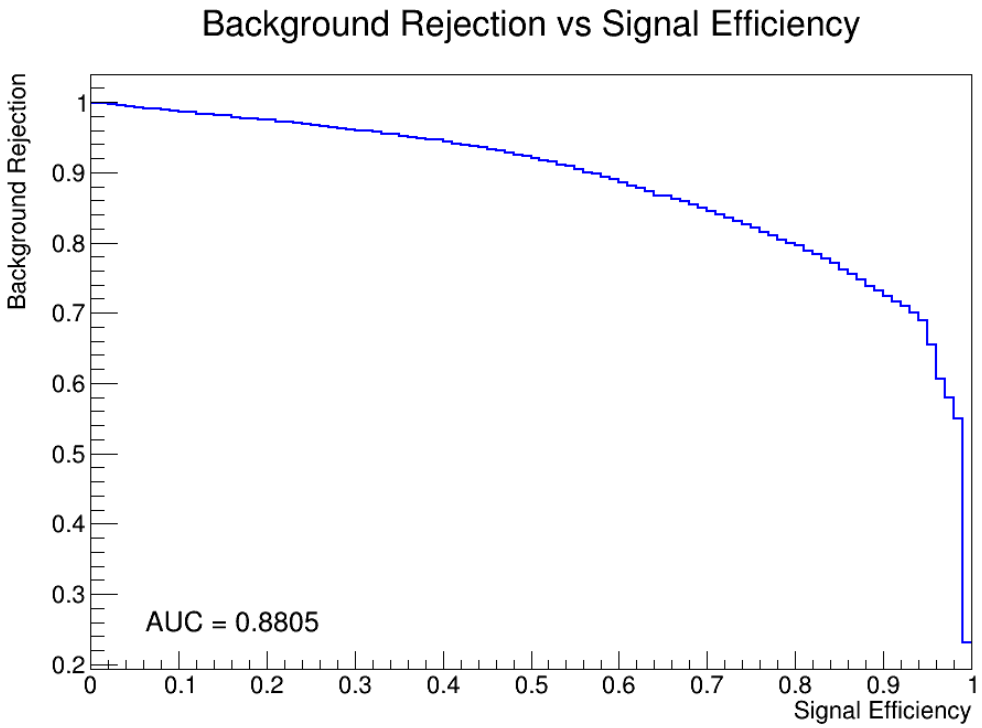

Рисунок 3.34 — ROC-кривая классификатора фиксированных отборов с 5 отобранными наблюдаемыми для этапа 3, отборы на наблюдаемые при 80% эффективности: $E_{\text{ratio}} > 0.564$, $w_{\eta 2} > 0$, $R_\phi > 0.254$, $\Delta E < 0.142$, $(M_2 + S_1)/S_9 < 4.71$

ROC-кривая для этапа 2, полученная аналогично, приведена на рисунке 3.35. Значение ROC AUC классификатора фиксированных отборов с отобранными наблюдаемыми составило 0.908 для этапа 2 и 0.881 для этапа 3.

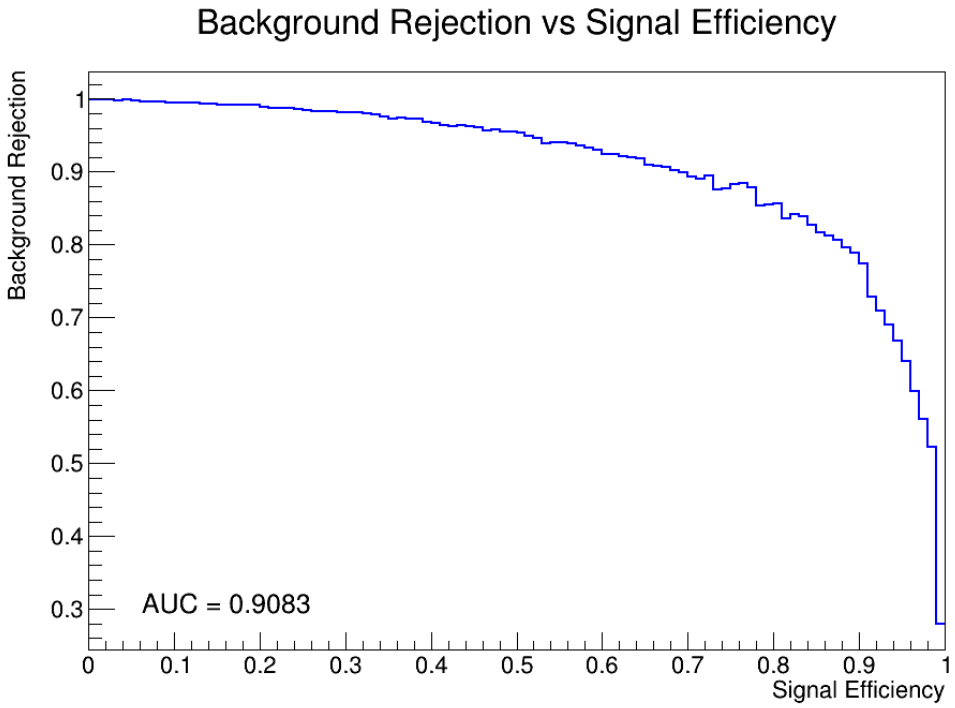

Рисунок 3.35 — ROC-кривая классификатора фиксированных отборов с 5 отобранными наблюдаемыми для этапа 2, отборы на наблюдаемые при 80% эффективности: $\Delta E < 0.056$, $S_6/S_9 > 0.963$, $S_{XY} > -0.00187$, $r^2 < 0.00098$, $TailImportance > 0.0670$

3.5 УЛУЧШЕНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ С ПОМОЩЬЮ КАТЕГОРИЗАЦИИ

Дополнительным способом повышения качества классификации является категоризация событий по положению кластера в калориметре. Как было описано ранее, калориметр SPD состоит из цилиндрической части и двух торцевых эндкапов, геометрия которых существенно различается. Форма электромагнитного ливня в цилиндрической части и эндкапах отличается из-за разной ориентации ячеек относительно направления частицы, поэтому разделение обучающей выборки на две области позволяет обучить специализированные классификаторы, лучше адаптированные к особенностям каждой из них.

Выборка разделялась по значению η_{central}: события с $|\eta| < 1.23$ относились к цилиндрической части, с $|\eta| \geq 1.23$ — к эндкапам. Для каждой области был обучен отдельный классификатор фиксированных отборов. При тестировании итоговая ROC-кривая строилась следующим образом: из XML-

файлов весов каждого классификатора извлекались точки их ROC-кривых — 100 пар $(\varepsilon_S, \varepsilon_B)$. Затем для каждой возможной пары точек из классификатора для цилиндрической части и эндкапного классификатора вычислялись комбинированные эффективности как среднее взвешенное по числу событий в соответствующей области:

$$\varepsilon_S^{\text{comb}} = \frac{N_S^{\text{bar}} \cdot \varepsilon_S^{\text{bar}} + N_S^{\text{end}} \cdot \varepsilon_S^{\text{end}}}{N_S^{\text{bar}} + N_S^{\text{end}}},$$

$$\varepsilon_B^{\text{comb}} = \frac{N_B^{\text{bar}} \cdot \varepsilon_B^{\text{bar}} + N_B^{\text{end}} \cdot \varepsilon_B^{\text{end}}}{N_B^{\text{bar}} + N_B^{\text{end}}}.$$

Для каждого значения сигнальной эффективности выбиралось максимальное подавление фона $1 - \varepsilon_B^{\text{comb}}$ среди всех комбинаций. По полученной кривой вычислялась площадь под кривой (ROC AUC).

Результат применения данного подхода на данных этапа 3 (minbias) представлен на рисунке 3.36:

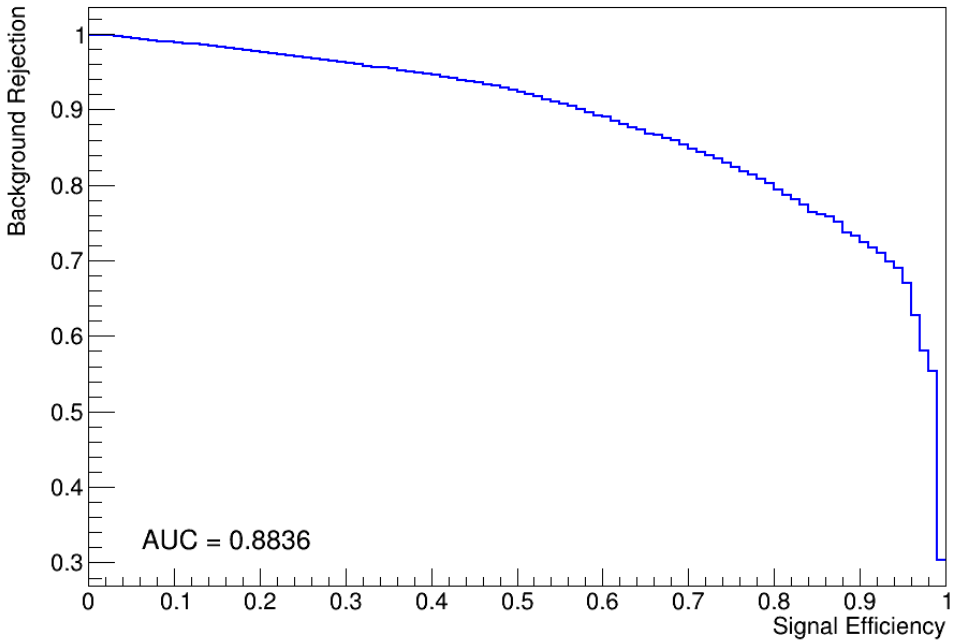

Рисунок 3.36 — ROC-кривая комбинированного классификатора фиксированных отборов с категоризацией по η на цилиндрическую часть и эндкапы для этапа 3

Применение описанной категоризации обеспечивает увеличение метрики ROC AUC классификатора фиксированных отборов приблизительно на 0.3%. Для BDT-классификатора аналогичная процедура не применялась, по-

скольку признак η_{central} уже был включён в набор обучающих наблюдаемых.

3.6 ИТОГОВОЕ СРАВНЕНИЕ КЛАССИФИКАТОРОВ

Итоговое сравнение метрик различных моделей на различных этапах тестирования представлено в таблице ниже:

Таблица 3.1 — Сравнение метрик классификаторов

Метрика	MLP	BDT (22 н.)	BDT (5 н.)	Cut-based
π^0 режекция при 80% γ эффективности (равномерное распределение по телесному углу, различные энергии)	90%	96%	—	88%
ROC AUC (равномерное распределение по телесному углу, различные энергии)	—	0.973	—	0.908
ROC AUC: Minbias, $E_{\text{cluster}} > 1\,\text{ГэВ}$	—	0.906	0.902	0.881
π^0 режекция при 80% γ эффективности (Minbias, $E_{\text{cluster}} > 1\,\text{ГэВ}$)	—	84%	83%	80%
π^0 режекция при 90% γ эффективности (Minbias, $E_{\text{cluster}} > 1\,\text{ГэВ}$)	—	78%	77%	74%

Таким образом, в ближайших условиях BDT-классификатор демонстрирует наилучшее качество разделения по сравнению как с MLP-классификатором, так и с классификатором фиксированных отборов. Следует, однако, отметить, что сопоставление с MLP-классификатором не является в полной мере прямым, несмотря на максимально возможное приближение условий: в оригинальном исследовании MLP обучался на выборке, превосходящей по объёму в 2.5 раза выборку из этапа 2, при этом распределение по энергиям было дискретным (шаг 0.5 ГэВ) в отличие от непрерывного в настоящей работе [6]. Дополнительным фактором, способным повлиять на результаты, является неизвестный из оригинального исследования баланс выборки (соотношение сигнальных и фоновых событий).

ЗАКЛЮЧЕНИЕ

В данной работе разрабатывались и тестировались алгоритмы идентификации фотонов в электромагнитном калориметре SPD.

Основные результаты:

- Адаптированы наблюдаемые фотонной идентификации из эксперимента ATLAS к геометрии калориметра SPD и реализованы в анализе.
- Изучены и адаптированы наблюдаемые, использованные ранее для решения данной задачи MLP-подходом, для совместного использования с наблюдаемыми ATLAS.
- Произведено ранжирование и отбор лучших параметров кластеров для различных классификаторов.
- Выполнено сравнение методов классификации кластеров γ/π^0: метод фиксированных отборов, MLP и BDT в различных условиях и сценариях.
- Показано, что BDT с расширенным набором параметров (ATLAS + SPDROOT) улучшает разделение при одинаковых условиях.
- На Minbias достигнуто лучшее значение ROC AUC 0.906 (BDT с 22 наблюдаемыми); BDT с 5 наблюдаемыми показал сопоставимый результат — 0.902.
- Определены две рабочие точки классификаторов с эффективностями идентификации фотонов 80% и 90%. На Minbias коэффициенты подавления фона составили: для BDT (22 н.) — 84% и 78%, для BDT (5 н.) — 83% и 77%, для классификатора фиксированных отборов — 80% и 74%.
- Определены значения фиксированных порогов для обеих рабочих точек. На основе этих данных можно начинать внедрение данного классификатора в среду SPDROOT.

- Показано, что категоризация по псевдобыстроте η обеспечивает прирост ROC AUC классификатора фиксированных отборов приблизительно на 0.3%.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Spin Physics Detector Project. — URL: https://spd.jinr.ru/. — (дата обр.: 28.01.2026).
2. Conceptual design of the Spin Physics Detector / V. M. Abazov [et al.]. — 2021. — arXiv: 2102.00442 [hep-ex]. — (дата обр.: 28.01.2026).
3. Probing Gluons with the Future Spin Physics Detector / A. Guskov [et al.] // Physics. — 2023. — Vol. 5. — P. 672–687. — (дата обр.: 28.01.2026).
4. Technical Design Report of the Spin Physics Detector at NICA / V. Abazov [et al.]. — 2024. — arXiv: 2404.08317 [hep-ex]. — (дата обр.: 28.01.2026).
5. SPDROOT: Software Framework for the SPD Experiment. — URL: https://git.jinr.ru/spd/spdroot.
6. *Maltsev A.* Separation of photon clusters from neutral pion decay in ECAL: exploratory study. — URL: https://indico.jinr.ru/event/5549/contributions/33072/attachments/23155/40918/SPD_PMC_20250903_Maltsev_ECAL_separation-1.pdf. — (дата обр.: 28.01.2026).
7. *Maltsev A.* Status of reconstruction in ECAL. — URL: https://indico.jinr.ru/event/5532/contributions/33735/attachments/23799/41947/Maltsev_SPD_CM_2025.pdf. — (дата обр.: 28.01.2026).
8. Toolkit for Multivariate Data Analysis with ROOT (TMVA) / K. Albertsson [et al.]. — URL: https://root.cern.ch/download/doc/tmva/TMVAUsersGuide.pdf. — (дата обр.: 28.01.2026).
9. *Goodfellow I.*, *Bengio Y.*, *Courville A.* Deep Learning. — MIT Press, 2016. — 781 p.

10. *ATLAS Collaboration*. Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016 // The European Physical Journal C. — 2019. — Vol. 79. — P. 205.

11. An introduction to PYTHIA 8.2 / T. Sjöstrand [et al.] // Computer Physics Communications. — 2015. — Vol. 191. — P. 159–177.