ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет Санкт-Петербургская школа физико-математических и компьютерных наук

Цобенко Маргарита Александровна

ПОИСК ФОТОНОВ СВЕРХВЫСОКИХ ЭНЕРГИЙ В ДАННЫХ ДЕТЕКТОРА KASCADE С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

Выпускная квалификационная работа
по направлению подготовки 01.04.02 Прикладная математика и информатика
образовательная программа «Машинное обучение и анализ данных»

Рецензент
ML Scientist, PicsArt, Inc.
_____
А.А. Котов

Научный руководитель
к. ф-м. н., доцент
_____
Д.Н. Москвин

Консультант
Старший лаборант в Институте Ядерной физики им. Будкера
_____
Н.А. Петров

Санкт-Петербург 2023

Оглавление

Введение

Космические лучи представляют собой поток заряженных частиц [1, 2]. К ним относятся высокоэнергетические ядра и элементарные заряженные частицы, такие как электроны и позитроны.

Космические лучи играют решающую роль в формировании химического состава межзвездной среды, нагреве межзвездного газа и определении динамики молекулярных облаков [3]. Это, в свою очередь, влияет на звездообразование и эволюцию галактик [4]. Космические лучи, в основном генерируемые взрывами сверхновых и образованием черных дыр, вызывают диффузные электромагнитные излучения вблизи своих источников и при распространении в межзвездной среде [2, 5]. Кроме того, космические лучи несут информацию о нестационарных, нетепловых и высокоэнергетических процессах во Вселенной [5]. Происхождение и точная природа космических лучей, особенно самых высоких энергий, до сих пор полностью не выяснены, что обусловлено сложностью процессов их образования [6].

Попадая в атмосферу Земли, первичные космические лучи, то есть внегалактические, галактические и солнечные космические лучи, взаимодействуют с ядрами азота, кислорода и генерируют каскад из миллионов вторичных частиц (протонов, фотонов, электронов, мезонов и других), которые частично достигают поверхности Земли и могут быть зарегистрированы приборами. Это явление называется широким атмосферным ливнем (Extensive Air Shower, EAS) [7]. Схематичное изображение приведено на Рисунке 1. Атмосферный ливень имеет три характерных компоненты: электромагнитную, мюонную и адронную [2].

Энергии вторичных частиц на несколько порядков меньше энергии исходной частицы. Эти частицы распространяются дальше и создают еще больше частиц при последовательных столкновениях с ядрами атмосферы.

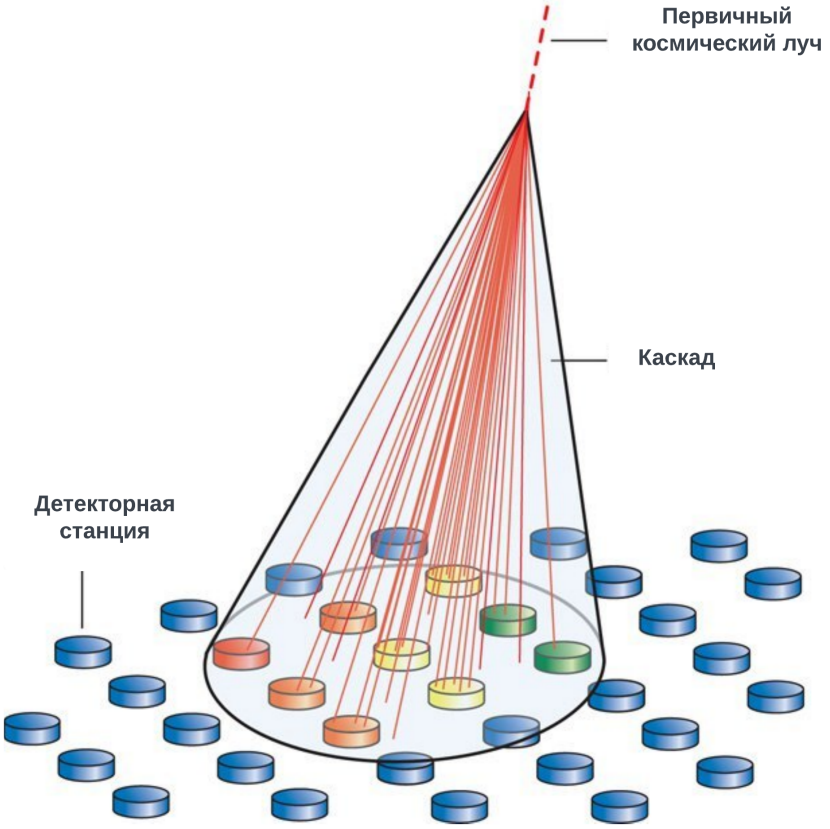


Рис. 1. Каскад вторичных частиц. Разные цвета обозначают разное время прибытия частиц [8].

С помощью набора детекторов вторичные частицы могут быть обнаружены и параметры падающей первичной частицы могут быть восстановлены. Существует множество детекторных установок по обнаружению космических лучей, например, *IceACT*, *LHAASO*, *TAIGA* и другие [9].

Широкие атмосферные ливни могут рождаться в результате прихода различных первичных частиц, например, адронов (класс составных частиц, подверженных сильному взаимодействию), а также фотонов. В данной работе будут рассмотрены атмосферные ливни, которые рождены гамма-квантами сверхвысоких энергий, то есть фотонами с энергией выше 10^{15} Эв.

Существуют исследования [10], в которых зарегистрированы гамма-кванты сверхвысоких энергий, обнаружено место, откуда они пришли, но проблема в том, что часть гамма-квантов приходит с направления, на кото-

ром нет мощного источника, например, черной дыры. В итоге ставится под сомнение механизм их ускорения [11]. Исследование гамма-квантов сверхвысоких энергий должно помочь понять свойства соответствующих источников излучения, механизмы происхождения космических лучей сверхвысоких энергий и свойства диффузного фона гамма-излучения. Узнав о происхождении этих частиц, человечество приблизится к созданию новых источников сверхвысоких энергий. Например, самый мощный ускоритель частиц на Земле, Большой адронный коллайдер, в разы уступает по энергии ускорения частиц космическим ускорителям [12]. Несмотря на то, что конкретные процессы в центре Галактики остаются неизвестными, ясно, например, что центр Галактики является потенциальным кандидатом на роль источника космических лучей с энергией ПэВ [13].

Задача по поиску гамма-квантов имеет ряд особенностей, которые усложняют поиск данных частиц. Первая проблема — наблюдательные сигнатуры (уникальные следы), оставляемые протонами и фотонами, имеют схожие характеристики. Поэтому данный тип частиц сложно выделить на фоне космических протонов. Вторая проблема — несбалансированность данных. Поток гамма-квантов очень мал, количество протонов в разы больше.

Текущее десятилетие положило начало новой области астрономии — гамма-астрономии сверхвысоких энергий. В последние годы в этой области наблюдается огромный прогресс как в теории, так и в наблюдениях, особенно в идентификации ПэВатронов (ускорителей космических лучей до энергии ПэВ, 10^{15} эВ).

В 2021 году обсерватория *LHAASO* опубликовала статью о наблюдении гамма-лучей сверхвысоких энергий (потока гамма-квантов) [10], было открыто множество кандидатов в ПэВатроны. В том же году *Tibet-AS*γ зарегистрировал диффузные гамма-кванты из галактической плоскости [14].

Данная работа посвящена исследованию данных *KASCADE* для дальнейшего их использования. Детектор *KASCADE* проработал порядка 15 лет, его данные опубликованы в открытом доступе [15]. Архив данных включает в себя порядка полумиллиарда зарегистрированных атмосферных ливней, а полная экспозиция составляет примерно половину экспозиции *LHAASO*, использованной в её исследовании [10]. Таким образом, данные *KASCADE* должны содержать события, вызванные первичными гамма-лучами с энергиями от нескольких сотен ТэВ до ПэВ, и задача сводится к тому, чтобы выделить эти события из гигантского фона протонных событий.

Методы машинного обучения успешно применяются в детекторном анализе частиц. В 2019 году от *LHAASO* вышла статья [16], в которой авторы используют градиентный бустинг над решающими деревьями для определения протонов. В 2020 году ученые в *LHAASO* для классификации протонов и не протонов использовали графовые нейронные сети [17]. Также в *LHAASO* в 2021 году уже для поиска гамма-квантов среди протонов использовали различные методы глубокого обучения [18]. В одной из последних работ, в которой используются данные *KASCADE*, для детекции гамма-частиц применяется алгоритм случайного леса [19]. Авторы смогли достичь уровня подавления фона порядка $10^{-3} - 10^{-2}$.

Целью данной работы является разработка методов классификации типа первичной частицы, которые позволили бы достичь сравнимого или более сильного уровня подавления фона, чем 10^{-3}. Для достижения обозначенной цели были определены следующие задачи:

- ознакомиться с предметной областью и существующими подходами;
- исследовать данные детектора *KASCADE*;
- построить модели классификации типа первичной частицы (гамма-

квант или протон);

- оценить качество предложенных методов на симуляционных данных детектора *KASCADE*;
- применить лучший классификатор к реальным данным.

Текст работы организован следующим образом: в Главе 1 описаны существующие работы по определению типа первичной частицы, использующие методы машинного обучения. В Главе 2 приведены описание эксперимента *KASCADE* и структура симуляционных данных. В Главе 3 представлены модели машинного обучения, которые используются в данной работе в качестве классификатора типа первичной частицы. В Главе 4 приведены результаты применения моделей классификации к симуляционным данным, в заключительной Главе 5 подведены итоги работы, а также возможные варианты продолжения исследования.

Глава 1

Обзор литературы

Существующие работы по определению типа первичной частицы можно разделить на две основные группы по типу используемых методов:

- работы, использующие традиционные методы, основанные на физике [10, 11, 14, 20, 21];
- работы, использующие методы машинного обучения [16–19, 22].

В этой главе будет приведен обзор работ, которые используют методы машинного и глубокого обучения, а также рассмотрены методологии предложенных подходов и проведен краткий анализ результатов.

1.1. *KASCADE*

Kostunin et al. [19] в своей работе проводят анализ массового состава, основанный на архивных данных, полученных с 1998 по 2013 год, предоставленных *KASCADE* Cosmic ray Data Center (KCDC). Для анализа авторы применяют современные методы машинного обучения. В качестве данных для обучения используются симуляционные данные, предоставленные KCDC. Одним из результатов данной работы является поиск кандидатов в гамма-кванты в области энергий ПэВ.

Моделью определения типа первичной частицы является случайный лес [23, 24], который представляет собой ансамблевый метод машинного обучения, использующий наборы деревьев решений для различных подвыборок обучающих данных для повышения точности по сравнению с одним деревом решений. Из-за несбалансированности данных, гамма-квантов го-

раздо меньше протонов, подход бинарной классификации не обеспечивает существенного уровня обнаружения, поэтому авторы остановились на модели регрессии случайного леса из 1000 деревьев, которая возвращает вероятность принадлежности к классу реконструированного события. Это позволило выбрать такое пороговое значение для вероятности принадлежности частицы к гамма-квантам, при котором уровень подавления фона получился порядка $10^{-3} - 10^{-2}$. В заключении авторы указывают на то, что данный подход требует дальнейшего развития и улучшения.

1.2. *LHAASO*

Large High Altitude Air Shower Observatory (*LHAASO*) — это обсерватория по наблюдению за обширными атмосферными ливнями, вызванными гамма-лучами и космическими лучами [25], в которой проводится эксперимент с измерением спектра космических лучей и их компонентов в масштабе ПэВ [26]. Эксперимент состоит из следующих детекторных станций:

- Extensive Air Shower (EAS) Kilometer Square Array (KM2A),
- Closed packed water Cherenkov detector array (WCDA),
- Cherenkov telescope array (WFCTA).

LHAASO-KM2A занимает основную площадь и состоит из двух наборов детекторов. Первый набор включает в себя 5195 детекторов электромагнитных частиц (electromagnetic particle detectors, ED), а второй набор состоит из 1171 детектора мюонных частиц (underground water Cherenkov ranks for muon detectors, MD). В состав WCDA входят три водоема с глубиной около 4 метров.

Расположение каждой компоненты *LHAASO* показано на Рисунке 1.1, где красная и синяя точки обозначают детекторы KM2A-ED и KM2A-MD

соответственно. Детекторы ED разделены на две части: центральную часть с 4901 детектором и внешнее кольцо с 294 детекторами. Массив MD играет ключевую роль в выделении гамма-квантов на фоне ядер космических лучей, а также дает важную информацию для классификации групп космических лучей.

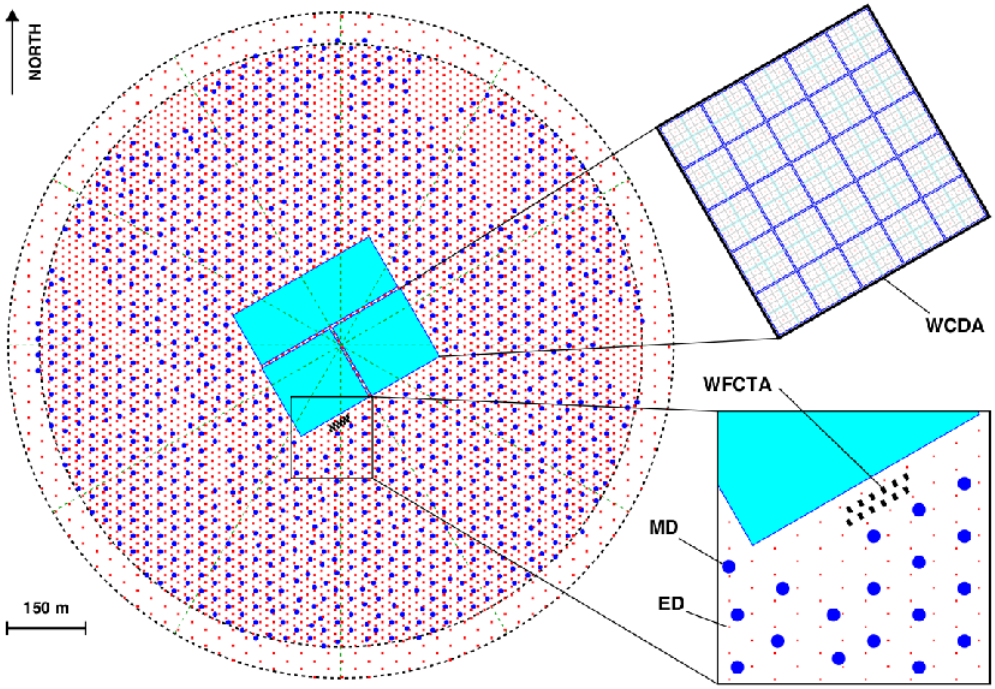


Рис. 1.1. Схема эксперимента *LHAASO*. На вставках показаны детали одного WCDA, а также ED (красные точки) и MD (синие точки) KM2A. Также показан WFCTA, расположенный на краю WCDA [17].

В работе [17] авторы используют графовую нейронную сеть (Graph Neural Network, GNN) для классификации компонентов космических лучей в эксперименте *LHAASO*-KM2A, где детектор, активированный событием, формируется в виде графа.

Детекторы ED и MD построены как взвешенные неориентированные плотные графы независимо друг от друга, каждый узел которых содержит трехмерный вектор признаков, в котором каждый признак нормируется независимо. Граф событий показан на Рисунке 1.2.

Матрица смежности A размера $n \times n$ определяется применением ядра

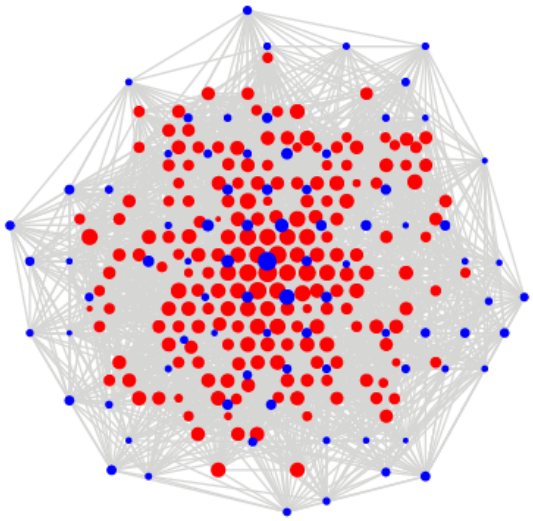


Рис. 1.2. Детекторы *LHAASO*-KM2A, структурированные в виде графа, активированные событием EAS с энергией 500 ТэВ, где красные точки представляют собой ED, а синие точки представляют MD [17].

Гаусса к попарному расстоянию $||x_i - x_j||$ между активированными детекторами:

$$d_{ij} = \exp(-\frac{1}{2}(||x_i - x_j|| - \mu_t)^2/\sigma_t^2),$$
$$a_{ij} = \frac{d_{ij}}{\sum_{k \in N} d_{ik}}. \tag{1.1}$$

В уравнении (1.1), a_{ij} представляет собой нормализованный элемент в матрице смежности, а N обозначает набор смежных детекторов по отношению к детектору i. μ_t и σ_t являются обучаемыми параметрами, которые определяют локальность ядра свертки. Кроме того, диагональные элементы в матрице A зануляются.

Сначала авторы извлекают многомерные признаки из входных векторов с помощью обучаемой функции, как показано в уравнении (1.2), из которого матрица вершин v размера $n \times 3$ преобразуется в матрицу $x^{(0)}$ размера $n \times d^{(0)}$.

$$x^{(0)} = ReLU(W^{(0)}v + b^{(0)}) \tag{1.2}$$

Затем авторы определяют последовательность слоев свертки размера T, как показано в уравнении (1.3). Каждый слой свертки t сначала агрегирует соседей путем умножения на матрицу смежности A и расширяет вектор от размерности $d^{(t)}$ до размерности $2 \cdot d^{(t)}$. Затем применяется весовая функция для обновления вектора до размера $d^{(t+1)}$. Используется нелинейная функция активации ReLU, за исключением последнего слоя свертки T.

$$GConv(x^{(t)}) = W^{(t)}[x^{(t)}, Ax^{(t)}] + b^{(t)}$$
$$x^{(t+1)} = \begin{cases} ReLU(GConv(x^{(t)})), \ t+1 < T \\ GConv(x^{(t)}), t+1 = T \end{cases} \tag{1.3}$$

Структура графа сохраняется при сверточных операциях. На последнем слое свертки, то есть на слое T, авторы добавляют global pooling layer, чтобы собрать признаки по всему графу и сжать граф в представление, не зависящее от размера. Матрица признаков $n \times d^{(T)}$ усредняется и преобразуется в $1 \times d^{(T)}$-мерную матрицу. Определение global pooling layer:

$$x_i^{(pool)} = \frac{1}{N} \sum_{n \in N} x_{ni}^{(T)} \tag{1.4}$$

На последнем уровне авторы используют линейный слой и применяют логистическую регрессию для классификации события

$$y = sigmoid(W^{(pool)} x^{(pool)} + b^{(pool)}), \tag{1.5}$$

где $x^{(pool)}$ — это $d^{(T)}$-мерный признак из global pooling слоя, а y — оценка классификации. Функция активации сигмоида обеспечивает принадлежность y диапазону $[0, 1]$, где сигналоподобное или фоноподобное событие приближается к 1 или 0 соответственно.

Авторы строят GNN для детекторов ED и MD независимо и объединяют их выходные данные вместе через линейный слой в уравнении (1.5),

где $x^{(pool)}$ представляет собой вектор размера $2 \cdot d^{(T)}$. По результатам работы модель GNN показала лучшие результаты, чем базовая физическая модель. В качестве метрики авторы использовали ROC-AUC. В этой статье также можно увидеть результаты работы модели CNN, результаты которой оказались хуже чем GNN.

В работе [18] для детекции протонов и гамма-квантов в LHAASO-KM2A авторы используют многослойный перцептрон и графовую нейронную сеть. Оба метода показывают хорошие результаты при детекции частиц во всех диапазонах энергий, используя метрику ROC-AUC.

Еще одной работой, в которой используются данные *LHAASO* является работа [16]. Сначала авторы проводят анализ параметров трех типов детекторов в соответствии с характеристиками EAS. И затем используют набор инструментов для многомерного анализа (Toolkit for Multivariate Analysis, TMVA) [27] для определения типа частицы.

Многофакторный анализ — важный раздел статистики, который применяется во множестве дисциплин. TMVA специально разработан для физики высоких энергий. Это инструмент для классификации сигналов и фона, который также позволяет идентифицировать компоненты космических лучей [28].

TMVA работает на основе машинного обучения. Он объединяет несколько расширенных алгоритмов-классификаторов, таких как Boosted Decision Trees (BDT), Artificial Neural Networks (ANN), метод опорных векторов (SVM) и так далее. Авторы в своей работе в качестве алгоритма классификации выбрали Boosted Decision Trees with Gradient boosting (BDTG). Авторы приводят исчерпывающую информацию о параметрах обучения и модели, но в итоге замечают, что результат отделения протона от других ядер едва ли удовлетворителен.

Глава 2

Эксперимент *KASCADE*

Эксперимент *KASCADE* [29] представляет собой детектор, состоящий из 252 детекторных станций, для измерения космических лучей высоких энергий посредством обнаружения широких атмосферных ливней. Основными компонентами детектора *KASCADE* являются массив *KASCADE*, детектор отслеживания мюонов (the Muon Tracking Detector) и центральный калориметр (the Central Calorimeter) для измерения адронной составляющей ливней. Мультидетекторные установки *KASCADE* и его расширение *KASCADE-Grande*, появившееся в 2003 году, прекратили активный сбор данных в 2013 году после более чем 20 лет сбора данных.

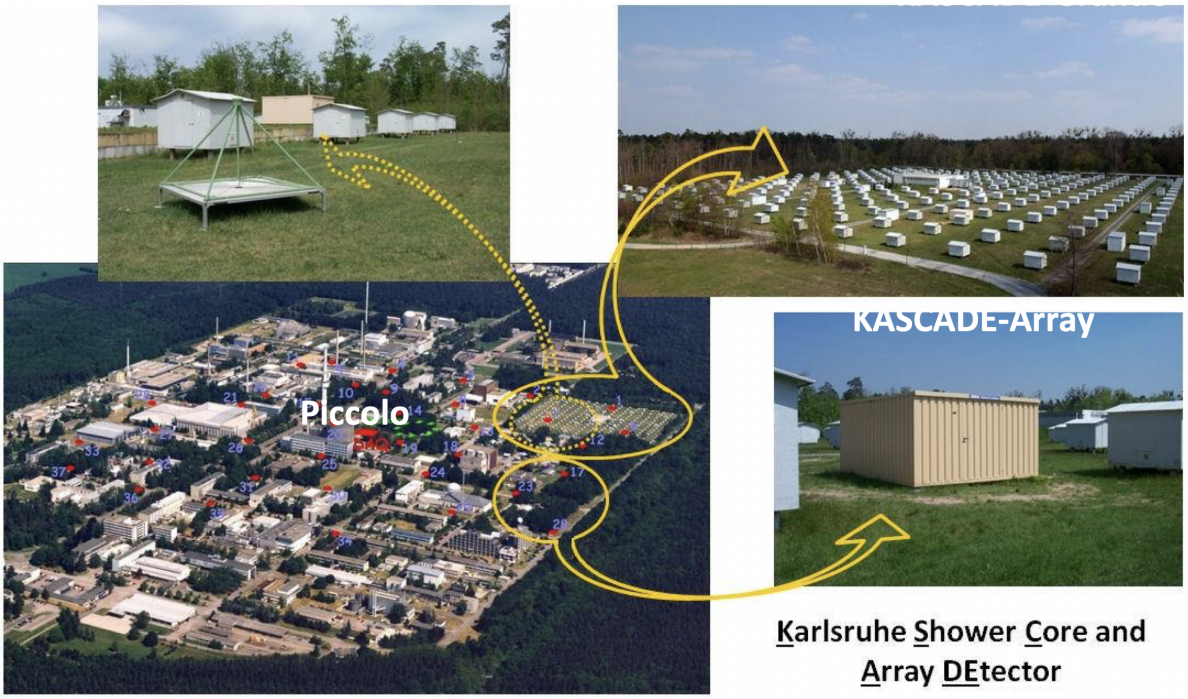


Рис. 2.1. Расположение детекторных станций KASCADE [30].

Данные предоставлены в открытом доступе (`https://kcdc.ikp.kit.edu`) и содержат измерения, которые включают в себя порядка полумиллиарда зарегистрированных атмосферных ливней. Центр данных космиче-

ских лучей *KASCADE* (KCDC) обеспечивает доступ к собранным данным о космических лучах экспериментов *KASCADE* и *KASCADE-Grande*.

2.1. Моделирование атмосферных ливней

Анализ экспериментальных данных по широким атмосферным ливням требует детального теоретического моделирования каскада, который развивается высокоэнергетической первичной частицей при входе в атмосферу. Этого можно достичь только с помощью подробных расчетов методом Монте-Карло с учетом всех знаний о сильных высокоэнергетических и электромагнитных взаимодействиях [29].

Так как поток частиц первичных космических лучей в диапазоне энергий, близких 1 ПэВ и выше, очень мал, в настоящее время возможны только косвенные измерения по регистрации широких атмосферных ливней, вызванных попаданием в атмосферу высокоэнергетических частиц. Поскольку первичные энергии ливней находятся за пределами энергетического диапазона искусственных ускорителей, а реакции, связанные с развитием ливней, происходят в направлениях, недоступных в коллайдерных экспериментах, неизбежны неопределенности в описании адронных взаимодействий в развитии ливней. Поэтому приходится полагаться на использование феноменологических моделей взаимодействия, которые в некоторых отношениях сильно различаются в своих предсказаниях, что еще более затрудняет задачу извлечения информации об отдельных энергетических спектрах из данных об атмосферных ливнях [29].

Моделирование атмосферных ливней для *KASCADE* представляет собой трехэтапную процедуру [29]:

1. моделирование атмосферного ливня, выполненное с помощью программы CORSIKA [31];

2. моделирование детектора, выполненное с помощью CRES;

3. реконструкция данных, выполненная с помощью KRETA.

Более подробную информацию можно найти в KCDC User Manual [30].

Программа CORSIKA (COsmic Ray SImulations for KAscade) позволяет моделировать взаимодействия и распады ядер, адронов, мюонов, электронов и фотонов в атмосфере до энергий порядка 10^{20} эВ. С помощью CORSIKA можно получить тип частицы, энергию, местоположение, направление и время прибытия всех вторичных частиц, которые создаются в атмосферном ливне и проходят выбранный уровень наблюдения.

Реализовано множество моделей адронных взаимодействий высоких и низких энергий. В *KASCADE* используют три разных семейства моделей: QGSJET, EPOS, SIBYLL [32]. Данные этих моделей доступны через веб-портал KCDC, чтобы пользователи могли выполнять свой собственный анализ массового состава. В данной работе CORSIKA используется для моделирования атмосферных ливней.

CRES (Cosmic Ray Event Simulation) представляет собой программный пакет для моделирования энерговыделения во всех компонентах детектора *KASCADE* / *KASCADE-Grande* в ответ на широкий атмосферный ливень, моделируемый с помощью CORSIKA. CRES принимает смоделированные в CORSIKA данные об атмосферных ливнях в качестве входных данных и возвращает смоделированные сигналы детектора.

Атмосферные ливни, измеренные *KASCADE*, анализируются с использованием программы реконструкции KRETA (KASCADE Reconstruction for ExTensive Air showers). Исходя из энерговыделений и индивидуальных временных меток во всех детекторах всех компонентов, KRETA определяет физические величины, такие как направление ливня или общее количество электронов, мюонов, адронов. KRETA считывает необработанные данные,

выполняет калибровку, реконструирует основные наблюдаемые ливни и сохраняет все результаты в виде гистограмм и векторов параметров.

В данной работе используются данные *KASCADE*, чтобы предсказать тип частицы (протон или гамма-квант). Данные моделирования берутся сразу для трех различных моделей адронных взаимодействий, а также для тестирования моделей, предложенных в данной работе, используются реальные данные. Признаки реконструируются с помощью традиционных эвристических подходов [29].

2.2. Структура данных

Данные детектора *KASCADE* структурированы в виде событий. Каждое событие описывается тремя матрицами размера 16×16. Они представляют собой время прихода частиц, электромагнитную и мюонную составляющие атмосферных ливней.

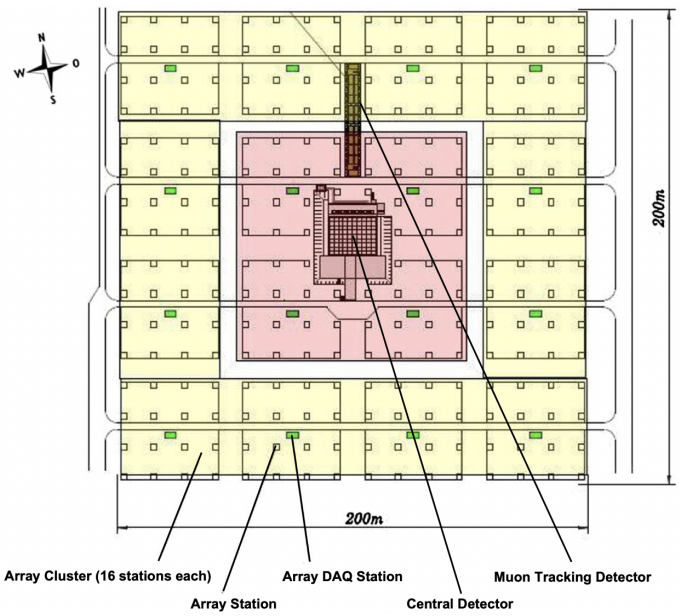


Рис. 2.2. Схематический вид детекторных станций *KASCADE* [30].

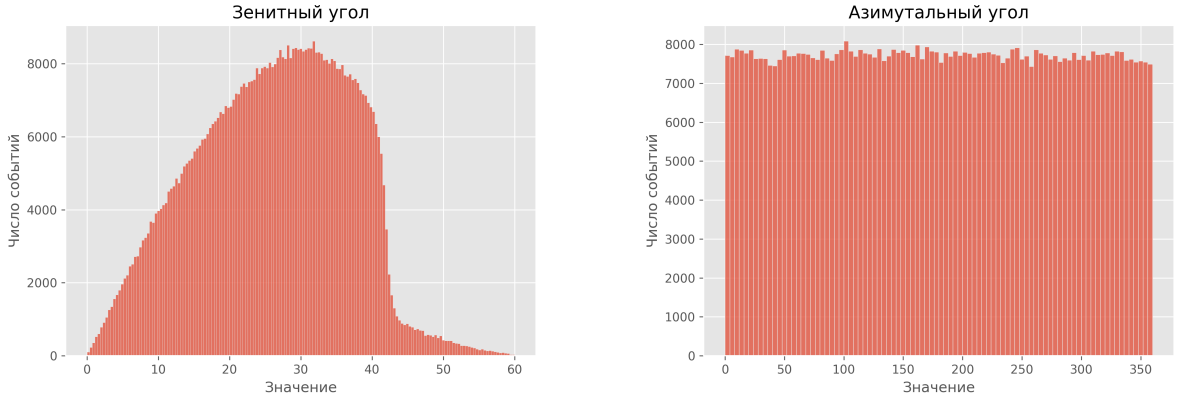
Также каждое событие описывается дополнительными реконструированными признаками, такими как энергия первичной частицы, направле-

ние прихода, количество электронов и мюонов и форма ливня. Ниже приведены краткие описания признаков.

Энергия первичной частицы (Primary Energy), Рисунок 2.3.

Рис. 2.3. Распределение значений энергии первичной частицы.

Зенитный угол измеряется относительно вертикального направления. Если его значение равно 0, то это указывает на вверх идущий ливень, а если значение равно 90 — это горизонтально идущий ливень. Азимутальный угол определяется как угол, измеренный по часовой стрелке, начиная с северного направления (90 градусов — восток), Рисунок 2.4.

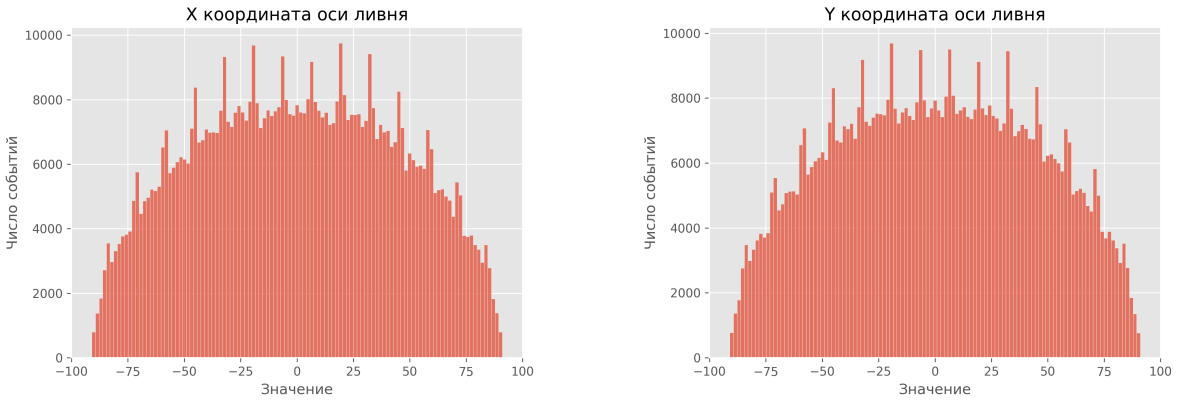


(*а*) (*б*)

Рис. 2.4. Распределение значений зенитного и азимутального углов.

За направление прихода отвечают следующие признаки, Рисунок 2.5:

- X-координата оси ливня (Core Position X) — расположение реконструированной оси ливня по оси x
- Y-координата оси ливня (Core Position Y) — расположение реконструированной оси ливня по оси y.



(*а*) (*б*)

Рис. 2.5. Распределение значений X и Y координат оси ливня.

Положение оси ливня от центра детектора (Core Distance), вычисляется на основе предыдущих двух признаков, Рисунок 2.6.

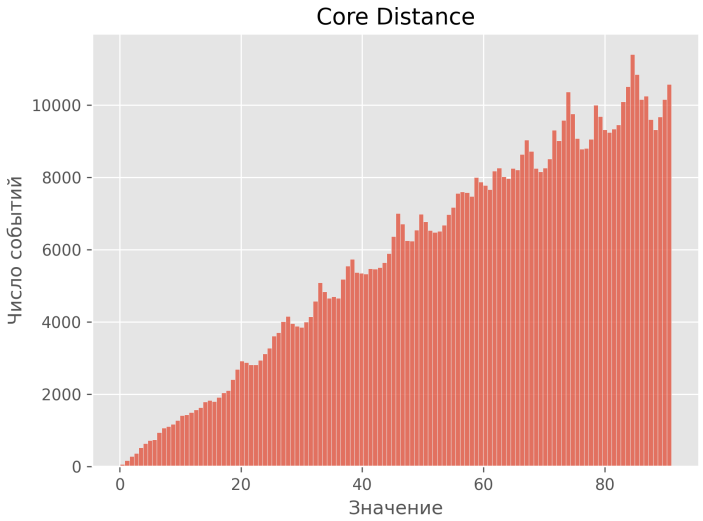
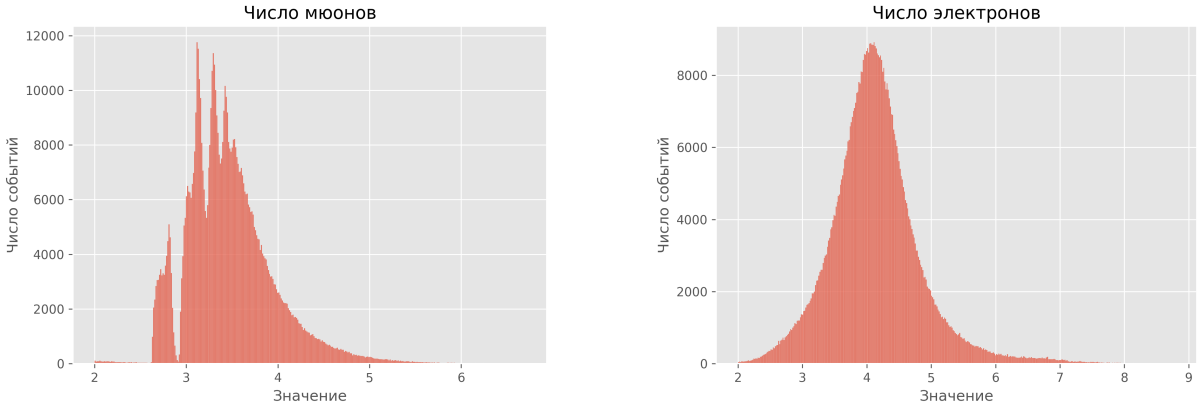


Рис. 2.6. Диапазон значений расстояний между осью ливня и центром детектора.



(*а*) (*б*)

Рис. 2.7. Диапазон значений числа мюонов и электронов.

Число мюонов и число электронов, Рисунок 2.7.

Параметр формы ливня (Age), Рисунок 2.8.

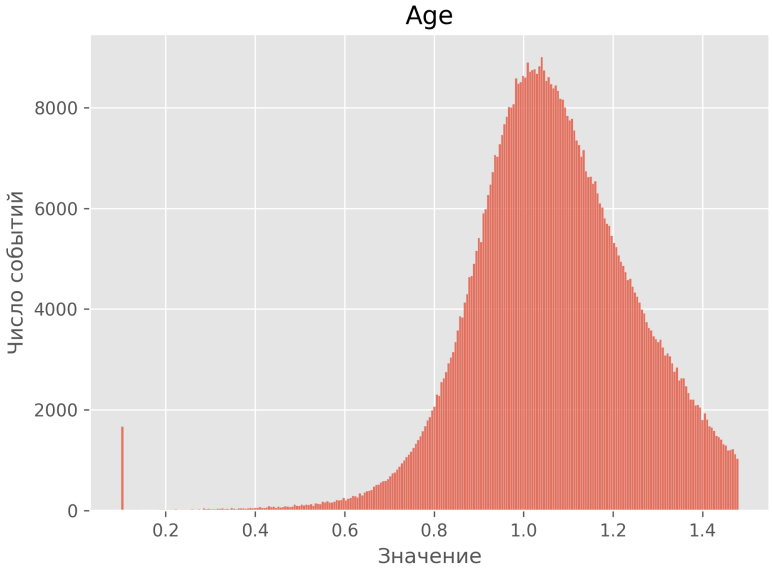


Рис. 2.8. Диапазон значений параметра формы ливня.

На Рисунках 2.9, 2.10 представлены примеры событий из реальных и смоделированных данных с близкими характеристиками энергии, зенитного угла и положения оси ливня от центра. В *KASCADE* электромагнитная и мюонная составляющие атмосферных ливней измеряются на различных детекторах.

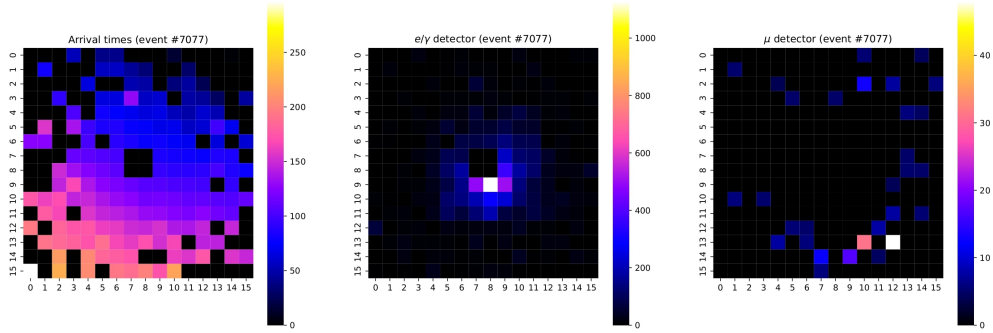


Рис. 2.9. Реальное событие.

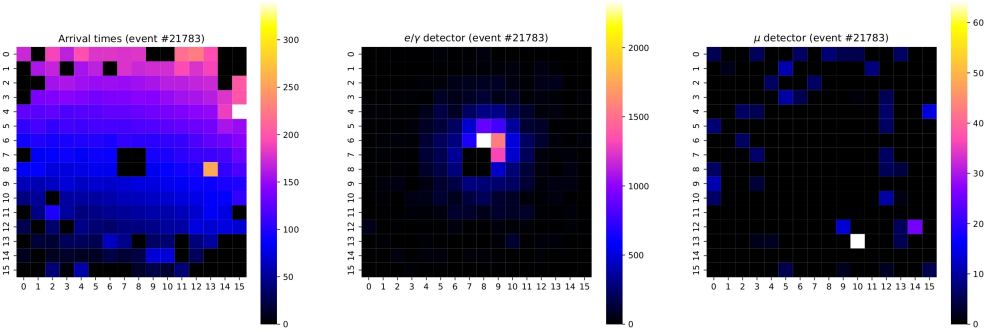


Рис. 2.10. Смоделированное событие.

На Рисунке 2.11 вычислены средние значения по всем матрицам. Черным цветом показаны участки, где детекторы отсутствуют.

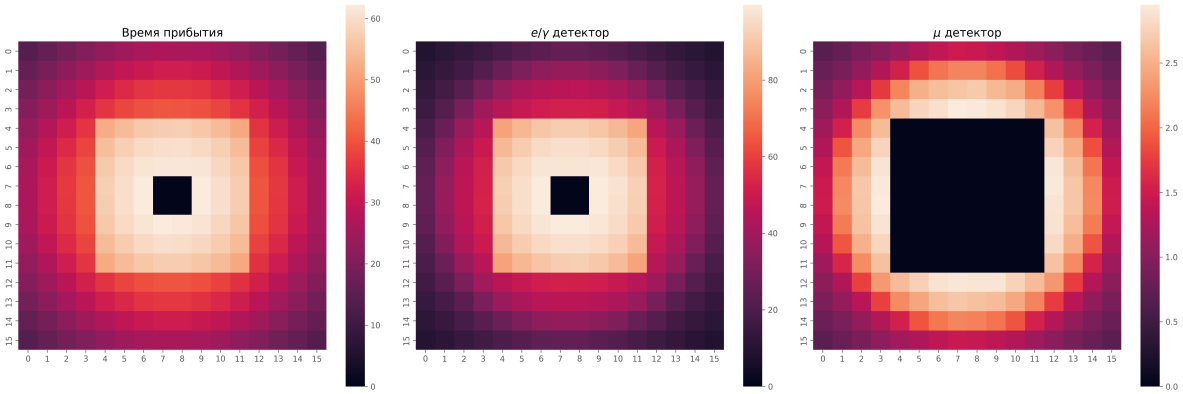


Рис. 2.11. Среднее по всем матрицам.

Глава 3

Модели классификации типа первичной частицы

В данной работе рассматриваются частицы только двух типов: протоны, которые представляют собой фон, и гамма-кванты, которые необходимо отделить от фона. Поэтому задачу детекции гамма-квантов можно представить в виде задачи бинарной классификации, где класс 0 — частица протон, класс 1 — целевой класс, частица гамма-квант. В данном разделе будет дано небольшое описание работы методов, которые были использованы в экспериментах по детекции гамма-частиц.

3.1. Логистическая регрессия

Пусть данные о частицах представлены в виде матрицы $X \in \mathbb{R}^{N\times(d+1)}$, то есть каждая частица представляет собой объект в некотором пространстве признаков, где первые d признаков реальные, а $d+1$ измерение соответствует константному признаку x_{d+1}, что позволяет не вводить дополнительно обозначение для сдвига w_0 и везде пользоваться нотацией скалярного произведения. Пусть также каждой частице соответствует одно из значений $Y = \{-1, 1\}$, где -1 будет означать, что частица является протоном, а 1 гамма-квантом. Тогда линейная модель классификации m определяется следующим образом:

$$\mathrm{m}(x) = \mathrm{sign}(\langle w, x\rangle),$$

где x — вектор признаков, а w — вектор весов, оба имеют размер $d + 1$.

Для обучения модели m ставится задача минимизации следующего

функционала:

$$Q(\mathrm{m}, X) = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{sign}(\langle w, x_i\rangle) \neq y_i), \tag{3.1}$$

то есть минимизации числа различий в предсказанном моделью классе и реальном классе частицы. Данный функционал (3.1) является дискретным относительно весов, поэтому принято минимизировать верхнюю оценку для $[y_i\langle w, x_i\rangle < 0]$. Если в качестве верхней оценки использовать $log(1 + \exp(-y\langle w, x\rangle))$, то линейную модель, обученную с таким функционалом, принято называть логистической регрессией [33].

3.2. Случайный лес

Случайный лес — это метод машинного обучения, который представляет собой бэггинг над решающими деревьями [23], то есть итоговая оценка представляет собой среднее базовых алгоритмов:

$$\mathrm{m}(x) = \frac{1}{K}\sum_{i=1}^{K} b_i(x).$$

При этом каждое дерево обучается на случайной подвыборке, полученной с помощью бутстрапа [23].

Алгоритм случайного леса эффективен при работе со сложными наборами данных, так как позволяет обрабатывать пропущенные значения. Данный алгоритм применяется в различных задачах. При этом качество его работы зависит от количества деревьев и настройки параметров, оптимизация которых может занять много времени.

3.3. Градиентный бустинг над решающими деревьями

Градиентный бустинг над решающими деревьями (Gradient Boosting Decision Trees, GBDT) — это алгоритм обучения с учителем, применяемый

в задачах, таких как регрессия и классификация [34]. Является представителем семейства ансамблевых методов обучения, то есть итоговый алгоритм представляет собой комбинацию базовых алгоритмов, что позволяет достичь лучшего качества работы. В GBDT базовые алгоритмы обучаются последовательно, каждое новое дерево решений обучается минимизировать ошибку предыдущего набора деревьев. Процесс последовательного обучения нескольких моделей для повышения точности итогового алгоритма называется бустингом. Таким образом, в результате обучения получается модель, прогнозом которой является объединение прогнозов всех деревьев. В качестве базовых алгоритмов чаще всего используются слабые, неглубокие деревья, так называемые пеньки.

Запишем более формально. Пусть имеется некоторая дифференцируемая функция потерь $L(y, \hat{y})$. Взвешенная сумма базовых алгоритмов выглядит следующим образом:

$$\mathrm{m}_N(x) = \sum_{n=0}^{N} k_n \cdot b_n(x).$$

В качестве начального алгоритма $b_0(x)$ чаще всего выбирают что-то простое, например, алгоритм, который возвращает только 0, среднее значение или самый популярный класс. Пусть построена композиция из $N-1$ алгоритма, тогда базовый алгоритм $b_N(x)$ выбирается так, чтобы минимизировать ошибку

$$\sum_{i=1}^{t} L(y_i, \mathrm{m}_{N-1}(x_i) + k_N \cdot b_N(x_i)) \to \min_{k_N, b_N}.$$

3.4. Многослойный перцептрон

Многослойный перцептрон (Multilayer Perceptron, MLP) — это многослойная нейронная сеть, состоящая из нескольких слоев взаимосвязанных

нейронов [35]. MLP применяется в задачах обучения с учителем, например, классификации и регрессии и является базовым блоком многих моделей глубокого обучения. Архитектуру MLP можно разделить на три основные части:

- входной слой, который принимает входные данные;
- один или несколько скрытых слоев, которые выполняют нелинейные преобразования;
- выходной слой, результатом которого является прогноз алгоритма.

Процесс обучения включает в себя минимизацию функции потерь. В процессе оптимизации корректируются веса нейронной сети с помощью обратного распространения ошибки.

Преимущество MLP состоит в том, что данная модель позволяет выучивать сложные нелинейные взаимосвязи между входными признаками и целевой переменной. Это делает многослойный перцептрон универсальной моделью для широкого круга задач.

3.5. Сверточная нейронная сеть

Сверточные нейронные сети [36] используются в различных задачах глубокого обучения. Одно из первых применений сверток в нейронных сетях — задача классификации изображений. Кроме задачи классификации изображений свертки используются в архитектурах сетей, которые решают задачу детекции объектов на изображении, сегментации изображений и анализе медицинских изображений, например, для определения раковых клеток. Важным свойством операции свертки является ее пространственная инвариантность (вариативность местоположения). Это означает,

что распознавание определенного объекта на изображении должно происходить вне зависимости от того, где именно на изображении расположен объект.

Как было сказано в Разделе 2, данные, получаемые с детекторных станций представлены в виде трех матриц размера 16×16. Таким образом, это позволяет представить их в виде трехканального изображения, где каждый «пиксель» кодируется тремя каналами: время прихода, электромагнитная и мюонная составляющие ливня. Архитектура сверточной сети, которая использовалась в экспериментах по детекции гамма-квантов, приведена на Рисунке 3.1.

На вход сети поступает тензор размера $3 \times 16 \times 16$, где первое измерение — число каналов, изначально оно равно 3. Второе и третье измерения отвечают за размер изображения, которое изначально имеет высоту и ширину равные 16. Сначала к входным данным последовательно применяются три свертки Conv_1, Conv_2, Conv_3 с последующими за каждой из сверток функциями активации ReLU. Параметры сверток следующие:

- Conv_1: число входных каналов равно 3, число выходных каналов равно 32,
  $\text{kernel size} = 3, \text{stride} = 1, \text{padding} = 0$;

- Conv_2: число входных каналов равно 32, число выходных каналов равно 64,
  $\text{kernel size} = 3, \text{stride} = 1, \text{padding} = 0$;

- Conv_3: число входных каналов равно 64, число выходных каналов равно 32,
  $\text{kernel size} = 3, \text{stride} = 1, \text{padding} = 0$.

В результате получается тензор размера $32 \times 10 \times 10$.

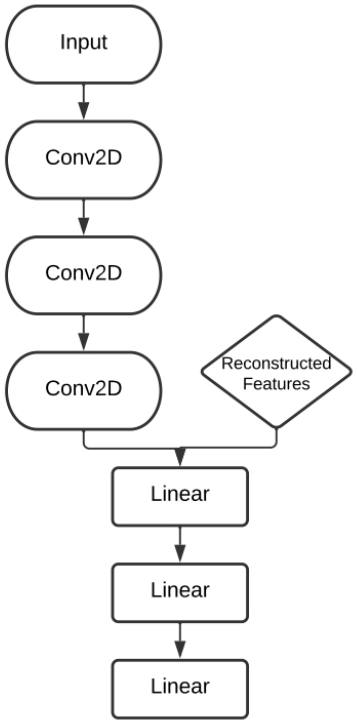


Рис. 3.1. Архитектура модели CNN.

Затем результат применения сверток представляется в виде вектора размера 3200. Данный вектор конкатенируется с вектором признаков, состоящим из реконструированных признаков события: энергия первичной частицы, направление прихода, количество электронов и мюонов, форма ливня. Полученный вектор проходит через многослойный перцептрон, состоящий из трех линейных слоев и использующий в качестве активаций на скрытых слоях ReLU. Число параметров получившейся сверточной сети равно 362117.

Результатом применения модели является логит, после применения к которому функции сигмоиды, получается вероятность того, что частица является гамма-квантом. Сравнивая данную вероятность с некоторым заданным пороговым значением, получается метка класса. Если вероятность больше порогового значения, то данная частица является гамма-квантом, в противном случае — это протон.

3.6. Графовая нейронная сеть

Графы являются одной из самых универсальных структур данных благодаря их огромной выразительной силе. Графовые нейронные сети [37] успешно используются в самых разных задачах машинного обучения, например, поиск возможных друзей в социальных сетях, предсказание побочных эффектов от лекарств [38], создание новых лекарств. Все это благодаря тому, что данные из реального мира часто можно представить в виде графов: в случае социальных сетей для каждого пользователя строится его социальный граф, позволяющий найти возможных друзей, предложить релевантные группы и сообщества. Графы, в которых каждое ребро между лекарствами может обозначать тип побочного эффекта, помогают определить возможные негативные последствия использования пары лекарств при лечении пациента.

Большим плюсом графового подхода к решению задач машинного обучения в отличие от рассмотренных ранее изображений, которые имеют строго заданные форму и размер, в том, что графы не обладают регулярной структурой. Это позволяет использовать подходы глубокого обучения в различных областях и учитывать нерегулярные особенности используемых данных. Один слой графовой нейронной сети — это fully-connected layer, веса которого применяются только к тем вершинам, которые являются соседями конкретной вершины в графе, в дополнение к ее собственному представлению с предыдущего слоя.

Каждое событие в *KASCADE* можно представить в виде взвешенного неориентированного графа, вершинами которого являются детекторные станции. Таким образом, задачу детекции частицы можно представить в виде задачи классификации графов — предсказать по графу, к какому классу (гамма-квант или протон) он принадлежит. Существует несколько спо-

собов задать значения длин ребер в графе. Первый — считать, что длины ребер между соседними вершинами (детекторными станциями) равны 1. Второй — в качестве длин ребер использовать реальные расстояния между детекторными станциями в поле. Каждая вершина графа будет представлена вектором признакового описания, который может быть получен из трех матриц, которыми описываются события (время прихода частиц, электромагнитная и мюонная составляющие атмосферного ливня). Архитектура графовой нейронной сети приведена на Рисунке 3.2.

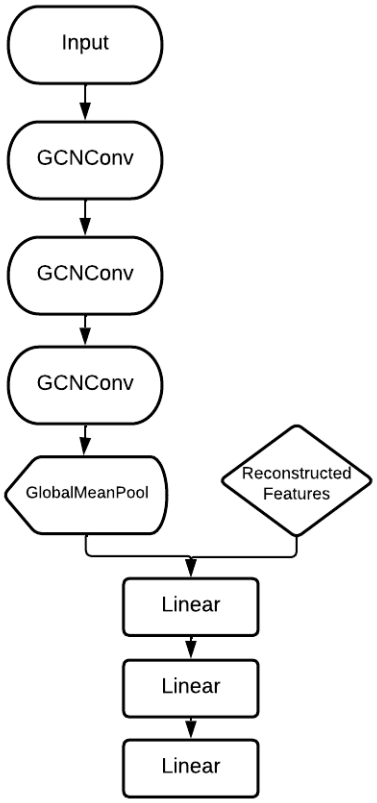


Рис. 3.2. Архитектура модели GCN.

Построенный граф события подается на вход графовой нейронной сети. Сначала граф проходит через 3 графовые свертки, предложенные Thomas Kipf и Max Welling [39], которые чередуются функциями активации ReLU. После этого применяется global mean pooling, который преобразовывает данный граф в вектор, который затем конкатенируется с вектором

признаков, описывающим каждое событие (энергия первичной частицы, направление прихода, количество электронов и мюонов и форма ливня). И далее вектор проходит через линейные слои, которые также чередуются функциями активации ReLU. Результатом работы модели является логит, который аналогично случаю в сверточной нейронной сети, можно перевести в вероятность принадлежности частицы к классу гамма-квантов.

Глава 4

Эксперименты

В данной главе будут описаны основные этапы проведения экспериментов по оценке качества работы предложенных в Разделе 3 алгоритмов классификации типа первичной частицы (гамма-квант или протон). Во всех экспериментах использовались симуляционные данные детектора *KASCADE*, структура которых описана в Разделе 2.

4.1. Подготовка данных

Симуляционные данные *KASCADE* представлены в виде событий, где каждое событие i описывается тензором M_i размера $3 \times 16 \times 16$ и вектором v_i, состоящим из 12 признаков. Из них 9 признаков — восстановленные признаки, а еще три дополнительных признака — это признаки, полученные с помощью зенитного и азимутального углов, преобразованием из сферических координат в декартовы. Также для каждого события определена метка $l_i \in \{0, 1\}$, где 1 означает, что частица является гамма-квантом, а 0 соответствует тому, что частица является протоном. Используемый набор данных состоит всего из 694810 событий, из которых 603562 событий являются протонами, то есть около 13.13% событий являются гамма-квантами.

Подготовка данных для обучения проводится следующим образом. Сначала все события случайно разбиваются на обучение, валидацию и тест в соотношении 54%, 23%, 23%. Затем для увеличения числа событий в обучающем наборе данных проводится аугментация. Аугментация данных внутри одного события осуществляется следующим образом: матрицы детекторов три раза поворачиваются на 90 градусов (Рисунок 4.1), при каж-

дом повороте осуществляется пересчет положения ядра ливня по X и по Y, а также азимутальный угол. После построения дополнительных событий случайным образом с вероятностью $p = 0.3$ выбираются примеры и добавляются в обучающий набор данных. В результате обучающий набор данных состоит из 792081 событий.

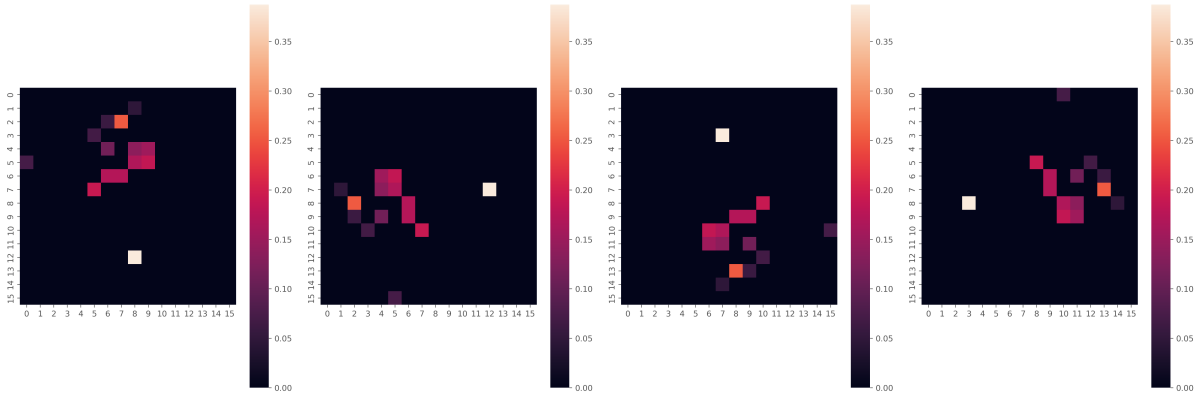


Рис. 4.1. Результаты поворота матрицы времени прихода при аугментации.

После этого проводится скейлинг данных. Всего используется три различных скейлера для матриц и один для признаков. Для матрицы времени прихода используется MinMaxScaler, для матриц энерговыделений электромагнитной составляющей и мюонной составляющей ливня, а также для вектора признаков используются три отдельных StandardScaler. Каждый скейлер обучается на данных с обучающего набора данных, после чего применяется к соответствующей компоненте каждого набора данных.

Каждая модель тренировалась и тестировалась на разных наборах данных различных адронных взаимодействий. Протоны были взяты из модели QGSJet для тренировочного и тестового наборов данных. Гамма-кванты были взяты из трех наборов данных: EPOS-LHC, SIBYLL, QGSJet и по-разному смешаны в тренировочном и тестовом наборах данных для увеличения статистики.

4.2. Метрики

Одним из важных показателей качества работы классификатора типа первичной частицы является уровень подавления фона. Подавление фона — это количество протонов, определенных как гамма-кванты, деленное на общее число протонов, что соответствует общепринятому понятию False Positive Rate. Исходя из результатов прошлых работ и поставленной цели, необходимо получить модели, которые будут иметь уровень подавления фона порядка $10^{-6} - 10^{-3}$.

Очевидно, что можно выбрать такое пороговое значение, что False Positive Rate будет равен 0, для этого достаточно выбрать $threshold = 1$. Именно поэтому второй важной характеристикой модели является доля выживших гамма-квантов, то есть отношение числа гамма-квантов, определенных как гамма-кванты, к общему числу гамма-квантов, что соответствует True Positive Rate, Recall.

Существует некоторый баланс между подавлением фона и числом выживших гамма-квантов. Поэтому при равном уровне подавления фона лучшей моделью будет та, у которой число выживших гамма-квантов больше.

Для выбора лучшей комбинации весов модели используется метрика ROC-AUC или площадь под ROC-кривой. Во время обучения после каждой эпохи осуществляется оценка метрики ROC-AUC на валидационном наборе данных и сохраняются веса модели. После обучения выбирается та версия модели (комбинация весов), при которой было получено максимальное значение метрики ROC-AUC на валидационном наборе данных.

4.3. Особенности обучения

В данной секции представлены основные аспекты обучения различных моделей классификации типа первичной частицы. Логистическая регрессия, XGBoost Classifier и MLP принимают на вход вектор размера 780, который состоит из спрямленных матриц и вектора восстановленных признаков. Модель случайного леса принимает на вход только вектор восстановленных признаков. Были проведены эксперименты, когда на вход подавался вектор размерности 780, но ограничение по вычислительным ресурсам не позволило построить ансамбль, который бы достиг лучшего качества, чем ансамбль, который использует только восстановленные признаки. Модели CNN и GNN принимают на вход тензор размера $3 \times 16 \times 16$, он проходит через свертки, затем результат представляется в виде вектора и конкатенируется с вектором признаков. Итоговый вектор идет на вход в голову классификации.

В качестве лосс-функции для моделей глубокого обучения используется бинарная кросс-энтропия. Обучение сети на несбалансированном наборе данных может привести к тому, что сеть будет склонна к предсказанию класса с большим количеством представителей, а другие классы будут игнорироваться. После того, как модель пройдет все эпохи обучения с бинарной кросс-энтропией в качестве лосса, берется лучшая комбинация весов модели и дополнительно обучается с Focal Loss [40]. Это позволяет компенсировать проблему несбалансированности данных. Focal Loss — это функция ошибки для обучения нейронных сетей, предложенная Lin et al. [40] для решения проблемы дисбаланса классов. Формула Focal Loss записывается следующим образом:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t),$$

где γ — гиперпараметр. Лучшее качество достигалось при дообучении с параметром $\gamma = 3$. Мотивация использовать Focal Loss для дообучения заключается в том, что это позволяет придать больше веса тем точкам, на которых модель все еще ошибается.

Модели Logistic Regression, Random Forest были реализованы с помощью библиотеки sklearn [41]. Для реализации градиентного бустинга над решающими деревьями использовалась библиотека XGBoost [42]. Нейронные сети (MLP, CNN, GNN) были реализованы с помощью библиотеки PyTorch [43]. Для оптимизации написания кода использовался фреймворк PyTorch Lightning [44]. Чтобы реализовать графовую структуру данных и графовые свертки использовалась библиотека PyTorch Geometric [45]. В качестве оптимизатора для обучения MLP, CNN, GNN использовался Adam [46] с шагом обучения $1e-3$. Для логирования метрик и прочей полезной информации использовался TensorBoard [47].

4.4. Результаты

В данной секции приведены результаты, полученные моделями на симуляционных данных *KASCADE*. В таблицах ниже представлены значения метрик, полученные в соответствующих диапазонах энергии и зенитного угла. В таблицах жирным шрифтом выделены лучшие результаты с точки зрения уровня подавления фона, полученные с помощью графовой нейронной сети. Эта модель достигает требуемого уровня подавления фона при энергиях > 16 и при энергиях от 15 до 16, когда зенитный угол от 20 до 40.

Для значений уровня подавления фона и доли выживших гамма-квантов были построены доверительные интервалы. В случае, где числа заменены обозначением $< 10^{-k}$, это означает, что правая граница доверительного

Таблица 4.1. Результаты моделей на обучении, валидации и тесте.

Модель	**Train**	**Validation**	**Test**		
			ROC-AUC	**TPR**	**FPR**
Logistic Regression	0.9346	0.9340	0.9335	0.4000	0.0051
Random Forest	0.9615	0.9382	0.9368	0.4564	0.0051
XGBoost	0.9482	0.9458	0.9448	0.5383	0.0051
MLP	0.9565	0.9486	0.9474	0.6061	0.0056
CNN	0.9609	0.9527	0.9518	0.6329	0.0053
GNN	0.9523	**0.9617**	**0.9607**	0.6139	**0.0004**

интервала оказалась меньше $< 10^{-k}$, например, при энергиях больше 16 правая граница доверительного интервала для значения метрики FPR для графовой нейронной сети меньше 10^{-3}. Тем самым подчеркивается, что был достигнут требуемый уровень подавления фона. Второй по качеству моделью является CNN.

В мануале указано, что при значении зенитного угла больше 42 и при $\log N_e \leq 4$, где N_e — число электронов, детектор *KASCADE* может работать некорректно (эффективность детектора низкая). Более детальную информацию можно найти в KCDC User Manual [30]. Поэтому в Таблице 4.2 приведены результаты с соответствующими ограничениями при разных диапазонах энергий.

В Таблицах 4.3, 4.4 приведены доверительные интервалы для значений метрик при различных значениях зенитного угла и энергии.

Таблица 4.2. Результаты моделей при различных значениях зенитного угла и энергии.

Модель	$0 \leq Ze \leq 20$ $15 \leq E \leq 16$		$20 < Ze \leq 40$ $15 \leq E \leq 16$		$0 \leq Ze \leq 20$ $E > 16$		$20 < Ze \leq 40$ $E > 16$	
	TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
Logistic Regression	0.8142	0.0051	0.7865	0.0051	0.8986	0.0041	0.9201	0.0086
Random Forest	0.7988	0.0059	0.7806	0.0054	0.9275	0.0061	0.9449	0.0060
XGBoost	0.8173	0.0055	0.8098	0.0054	0.9348	0.0061	0.9449	0.0215
MLP	0.7895	0.0018	0.7725	0.0011	0.9348	0.0051	0.9421	$< 10^{-3}$
CNN	0.7833	0.0007	0.7666	$< 10^{-3}$	0.9203	$< 10^{-3}$	0.9283	$< 10^{-3}$
GNN	0.7647	0.0004	0.7876	$\boxed{< \mathbf{10^{-4}}}$	0.9214	$\boxed{< \mathbf{10^{-3}}}$	0.9394	$\boxed{< \mathbf{10^{-3}}}$

Таблица 4.3. Доверительные интервалы значений метрик при различных значениях зенитного угла и энергии.

Модель	$0 \leq Ze \leq 20$ $15 \leq E \leq 16$		$20 < Ze \leq 40$ $15 \leq E \leq 16$	
	TPR	**FPR**	**TPR**	**FPR**
Logistic Regression	0.8144 ± 0.0437	0.0055 ± 0.002	0.7844 ± 0.0251	0.005 ± 0.0014
Random Forest	0.8025 ± 0.0533	0.0061 ± 0.0021	0.7819 ± 0.0257	0.0056 ± 0.0013
XGBoost	0.8218 ± 0.0464	0.0056 ± 0.0023	0.806 ± 0.0256	0.0053 ± 0.0014
MLP	0.7891 ± 0.0555	0.0018 ± 0.0011	0.773 ± 0.0247	0.0012 ± 0.0006
CNN	0.7842 ± 0.0388	0.0008 ± 0.0006	0.7663 ± 0.0265	$< 10^{-3}$
GNN	0.7616 ± 0.0473	0.00052 ± 0.00048	0.7904 ± 0.033	$< 10^{-4}$

Таблица 4.4. Доверительные интервалы значений метрик при различных значениях зенитного угла и энергии.

Модель	$0 \leq Ze \leq 20$, $E > 16$		$20 < Ze \leq 40$, $E > 16$	
	TPR	**FPR**	**TPR**	**FPR**
Logistic Regression	0.8926 ± 0.0563	0.005 ± 0.0049	0.9214 ± 0.0282	0.0093 ± 0.0051
Random Forest	0.9269 ± 0.0423	0.006 ± 0.0057	0.9429 ± 0.0253	0.0066 ± 0.0049
XGBoost	0.9328 ± 0.0453	0.0072 ± 0.007	0.9411 ± 0.02	0.0216 ± 0.0078
MLP	0.9304 ± 0.0378	0.007 ± 0.0069	0.9371 ± 0.0249	$< 10^{-3}$
CNN	0.9174 ± 0.0481	$< 10^{-3}$	0.9264 ± 0.0281	$< 10^{-3}$
GNN	0.9188 ± 0.0455	$< 10^{-3}$	0.9423 ± 0.0227	$< 10^{-3}$

Глава 5

Заключение

Целью данной работы была разработка методов классификации типа первичной частицы (гамма-квант или протон), которые позволили бы достичь сравнимого или более сильного уровня подавления фона, чем 10^{-3}. Данная цель была достигнута, также в результате работы было сделано следующее:

1. Реализованы модели классификации типа первичной частицы;
2. Проведена оценка качества предложенных методов на симуляционных данных детектора *KASCADE*;
3. Достигнут более сильный уровень подавления фона, чем 10^{-3};
4. Лучшая модель была применена к реальным данным;
5. Результаты работы были представлены:
    - Конференция High Energy Astrophysics (HEA), 2022, ссылка;
    - Workshop on Machine Learning for Cosmic-Ray Air Showers, Delaware university, 2022, ссылка;
    - Статья в рамках конференции European Cosmic Ray Symposium (ECRS), 2022, ссылка.
6. Реализация моделей доступна онлайн на GitHub.

В результате применения графовой нейронной сети к реальным данным были получены соизмеримые значения уровня подавления фона. Это

означает, что модель работает адекватно на реальных данных. Чтобы провести более тщательную проверку классификатора и улучшить качество моделей, требуется больше симуляций.

Дальнейшая работа также может быть направлена на использование классификатора для расширенного нерегулярного детектора *KASCADE-Grande*. Кроме того можно провести исследование абляции частиц высоких энергий с помощью имеющихся данных.

Список литературы

1. P. A. Zyla *et al.*, "Review of Particle Physics," *PTEP*, vol. 2020, no. 8, p. 083C01, 2020.
2. T. K. Gaisser, R. Engel, and E. Resconi, *Cosmic Rays and Particle Physics*, 2016.
3. N. Gnedin *et al.*, *Star Formation in Galaxy Evolution: Connecting Numerical Models to Reality: Saas-Fee Advanced Course 43. Swiss Society for Astrophysics and Astronomy*, Jan. 2016.
4. P. Hopkins *et al.*, "Effects of different cosmic ray transport models on galaxy formation," *Monthly Notices of the Royal Astronomical Society*, vol. 501, no. 3, pp. 3663–3669, Mar. 2021.
5. R. Aloisio, E. Coccia, and F. Vissani, *Multiple messengers and challenges in astroparticle physics*, Jan. 2018.
6. L. A. Anchordoqui, "Ultra-high-energy cosmic rays," *Phys. Rept.*, vol. 801, pp. 1–93, 2019.
7. P. K. F. Grieder, *Extensive Air Showers: High Energy Phenomena and Astrophysical Aspects - A Tutorial, Reference Manual and Data Book*, 2010.
8. P. Bauleo and J. Rodriguez Martino, "The dawn of the particle astronomy era in ultra-high-energy cosmic rays," *Nature*, vol. 458, pp. 847–51, 05 2009.
9. A. M. W. Mitchell, "Status of ground-based and galactic gamma-ray astronomy," 2021.
10. Z. Cao, F. Aharonian *et al.*, "Ultrahigh-energy photons up to 1.4 petaelectronvolts from 12 γ-ray galactic sources." *Nature*, 2021.
11. K. Kotera and A. V. Olinto, "The astrophysics of ultrahigh-energy cosmic rays," *Annual Review of Astronomy and Astrophysics*, vol. 49, no. 1, pp. 119–153, 2011. [Online]. Available: https://doi.org/10.1146/annurev-astro-081710-102620

12. CERN, "Accelerator report: Crescendo at the lhc following the first stable beams at 6.8 tev," 2023, https://home.cern/news/news/accelerators/accelerator-report-crescendo-lhc-following-first-stable-beams-68-tev [Accessed: May 2023].
13. H. collaboration A. Abramowski, F. Aharonian *et al.*, "Acceleration of petaelectronvolt protons in the galactic centre," *Nature*, vol. 531, pp. 476–479, 2016.
14. M. Amenomori, Y. Bao *et al.*, "First detection of sub-pev diffuse gamma rays from the galactic disk: Evidence for ubiquitous galactic cosmic rays beyond pev energies." *Physical review letters*, vol. 126 14, p. 141101, 2021.
15. A. Haungs, D. Kang *et al.*, "The kascade cosmic-ray data centre kcdc: granting open access to astroparticle physics research data," *The European Physical Journal C*, vol. 78, pp. 1–16, 2018.
16. L. Yin, S. S. Zhang *et al.*, "The expectation of cosmic ray proton and helium energy spectrum below 4 pev measured by lhaaso," *arXiv: High Energy Astrophysical Phenomena*, 2019.
17. C. Jin, S. zhan Chen, and H. He, *Classifying cosmic-ray proton and light groups in LHAASO-KM2A experiment with graph neural network*, 2020.
18. F. Zhang, "Identification of proton and gamma in lhaaso-km2a simulation data with deep learning algorithms," *Proceedings of 37th International Cosmic Ray Conference — PoS(ICRC2021)*, 2021.
19. D. Kostunin, I. V. Plokhikh *et al.*, "New insights from old cosmic rays: A novel analysis of archival kascade data," *Proceedings of 37th International Cosmic Ray Conference — PoS(ICRC2021)*, 2021.
20. L. Yin *et al.*, "Accurate measurement of the cosmic ray proton spectrum from 100tev to 10pev with lhaaso," 2017.
21. M. Amenomori, Y. Bao *et al.*, "First detection of photons with energy beyond 100 tev from an astrophysical source." *Physical review letters*, vol.

123 5, p. 051101, 2019.

22. R. U. Abbasi, M. Ackermann *et al.*, "Graph neural networks for low-energy event classification & reconstruction in icecube," *Journal of Instrumentation*, vol. 17, 2022.
23. L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
24. G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, pp. 197–227, 2015.
25. S. Vernetto, "Gamma ray astronomy with lhaaso," *Journal of Physics: Conference Series*, vol. 718, 2016.
26. H. He, "Design of the lhaaso detectors," *Radiation Detection Technology and Methods*, vol. 2, 06 2018.
27. A. Höcker, P. Speckmayeret *et al.*, "Tmva - toolkit for multivariate data analysis," 2007.
28. Z. Zong, B.-Y. Bi *et al.*, "Primary particle identification with mva method for the lhaaso project," 2017.
29. A. Haungs, D. Kang *et al.*, "The KASCADE cosmic-ray data centre KCDC: granting open access to astroparticle physics research data," *The European Physical Journal C*, vol. 78, no. 9, sep 2018.
30. KCDC, "Kcdc user manual," 2022, https://kcdc.ikp.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf [Accessed: May 2023].
31. D. Heck, T. Pierog, and J. Knapp, "Corsika: An air shower simulation program," 2012.
32. T. Pierog, *Review of Model Predictions for Extensive Air Showers*. [Online]. Available: https://journals.jps.jp/doi/abs/10.7566/JPSCP.19.011018
33. D. W. Hosmer and S. Lemeshow, "Applied logistic regression," 1991.
34. J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *Annals of Statistics*, vol. 29, pp. 1189–1232, 2001.
35. M. Popescu *et al.*, "Multilayer perceptron and neural networks," *WSEAS*

Transactions on Circuits and Systems archive, vol. 8, pp. 579–588, 2009.

36. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998.
37. F. Scarselli *et al.*, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.
38. M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, pp. i457 – i466, 2018.
39. T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv*, vol. abs/1609.02907, 2016.
40. T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2020.
41. F. Pedregosa, G. Varoquaux *et al.*, "Scikit-learn: Machine learning in python," *ArXiv*, vol. abs/1201.0490, 2011.
42. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
43. A. Paszke, S. Gross *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *ArXiv*, vol. abs/1912.01703, 2019.
44. "Pytorch lightning," 2019, https://www.pytorchlightning.ai/index.html.
45. M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *ArXiv*, vol. abs/1903.02428, 2019.
46. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
47. TensorFlow, "Tensorboard," 2015, https://www.tensorflow.org/tensorboard?hl=ru.