Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский ядерный университет МИФИ»
(НИЯУ МИФИ)

УДК 539.120.71

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

# Применение методов машинного обучения для лучшего выделения процессов рассеяния векторных бозонов

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

Консультант _____ А. М. Петухов

Студент _____ К. М. Савельев

Москва 2020

# Содержание

# Введение

Стандартная модель (СМ) даёт достаточно точные качественные и количественные предсказания для многих физических процессов. Однако, существует ряд явлений, которые не могут быть описаны в рамках Стандартной модели. Одни из самых известных: явление осцилляций нейтрино, из которого следует наличие у него массы, что противоречит СМ; тёмная материя, косвенные признаки наличия которой наблюдаются в большом масштабе в астрономических наблюдениях; факт барионной асимметрии. Всё это свидетельствует о том, что современная СМ не является всеобъемлющей. Это даёт стимул к поискам отклонений, которые могут привести к открытию более совершенной модели для описания взаимодействий элементарных частиц.

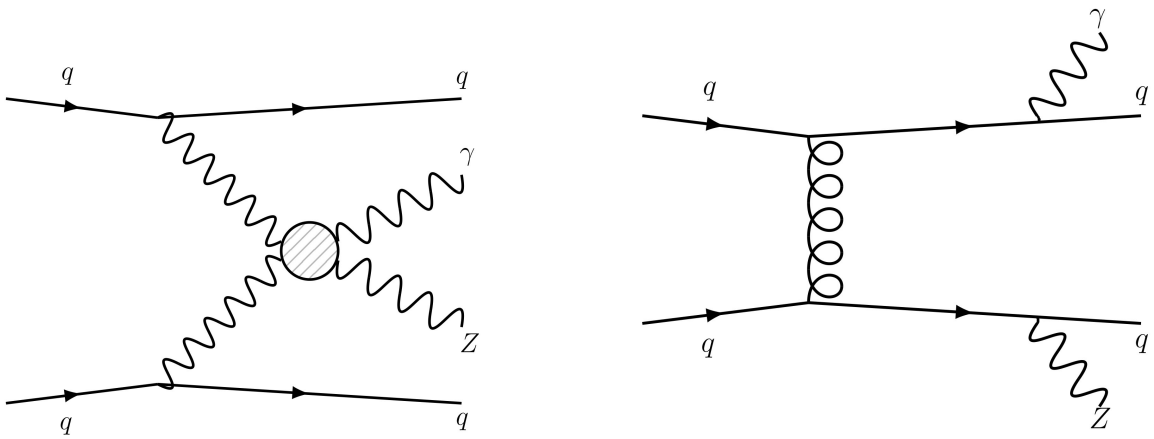
Целью исследования является поиск отклонений от предсказаний СМ при рассеянии векторных бозонов, $VV \to VV$, где $V = W/Z/\gamma$. Для исследования был выбран высокочувствительный к отклонениям СМ и на данный момент экспериментально не обнаруженный процесс электрослабого рождения $Z$-бозона, фотона совместно с двумя адронными струями с последующим распадом $Z$-бозона на нейтрино и антинейтрино. Выбор нейтрального канала распада связан с его достаточно большой вероятностью (20%) и возможностью отделения сигнала в отличии от распада по адронному каналу, вероятность которого составляет около 70%. Лептонный канал распада не рассматривался из-за его сравнительно низкой вероятности ($\sim 6.7\%$).

Этот процесс невозможно отделить от других электрослабых процессов с тем же конечным состоянием. Поэтому его изучение возможно только посредством рассмотрения всех процессов электрослабого образования конечного состояния $Z\gamma jj$. Они включают в себя как процессы рассеяния векторных бозонов, чувствительных к изменениям параметров Стандартной модели, так и прочие электрослабые процессы. Пример диаграммы процесса рассеяния представлен на рисунке 1a. Кинематические парамет-

ры частиц в конечном состоянии позволяют отделить их от КХД процессов с тем же конечным состоянием, которые являются основным фоном наряду с экспериментальными фонами. Пример диаграммы такого процесса представлен на рисунке 1б. Экспериментальные фоны включают в себя фоны, связаные с неверной идентификацией частиц, например, регистрацией электрона как фотона или струи как фотона, неправильное измерение потерянного поперечного импульса в событиях со струями.

Была выполнена проверка оправданности применения методов машинного обучения к отделению сигнальных событий от фона при изучении процессов электрослабого рождения $Z$-бозона, фотона и двух адронных струй, так как применение одномерных фиксированных отборов не даёт достаточной величины значимости, от которой зависит точность измерения сечения процесса.

В ходе работы было проведено сравнение максимальной величины значимости при использовании фиксированных отборов по переменным и применении классификатора. Эффективное разделение сигнальных и ых событий даст возможность определить величину сечения исследуемого процесса с большей точностью.




(а) (б)

Рисунок 1 — Диаграммы процессов (а) образования посредством рассеяния векторных бозонов и (б) КХД образование состояния $Z\gamma jj$

Результаты измерения параметров этого процесса может использоваться для поиска аномалий в вершине $WWZ\gamma$, а также вершин $ZZZ\gamma$,

$ZZ\gamma\gamma$ и $Z\gamma\gamma\gamma$, запрещённых в СМ.

В предыдущем исследовании, опубликованном коллаборацией ATLAS, использовались данные столкновений с энергией в системе центра масс 8 TeV и интегральной светимостью 20.3 $\text{fb}^{-1}$, из-за недостаточной чувствительности было выполнено только для поиска аномальных вершин [1]. В свою очередь, это исследование позволит достичь значимости измерения сечения процесса в $5\sigma$. Эта величина является необходимым минимумом для обнаружения этого процесса. На данный момент опубликовано сечение процесса $Z(\ell\ell)\gamma jj$ со более $3\sigma$ [2].

# 1 Методы

## 1.1 Алгоритм BDT

Алгоритм BDT (Boosted Decision Trees) [3] – это классификатор с бинарной древовидной структурой. Принцип его работы заключается в поочерёдном применении ограничений по различным переменным. Это позволяет разбить фазовое пространство на множество областей, которые классифицируются как сигнальные или фоновые.

Алгоритм был выбран из-за простоты применения, а также отсутствия необходимости в его настройке. Переменные для обучения не требуют никакой предварительной подготовки. Благодаря применению бустинга он не склонен к переобучению.

## 1.2 Принцип построения дерева решений

Входные данные попадают в корневой узел дерева, далее производятся отборы по переменным так, чтобы максимизировать коэффициент разделение сигнала и фона. Затем из этих отборов выбирается тот, который обеспечивает максимальное разделение событий. Процесс повторяется для каждого дочернего узла до тех пор, пока количество событий в какомлибо из них не станет меньше установленного. Далее все узлы классифицируются как сигналоподобные или фоноподобные в зависимости от коэффициента чистоты или от преобладания в них сигнальных, либо фоновых событий. Схематичный вид дерева решений изображен на рисунке 2.

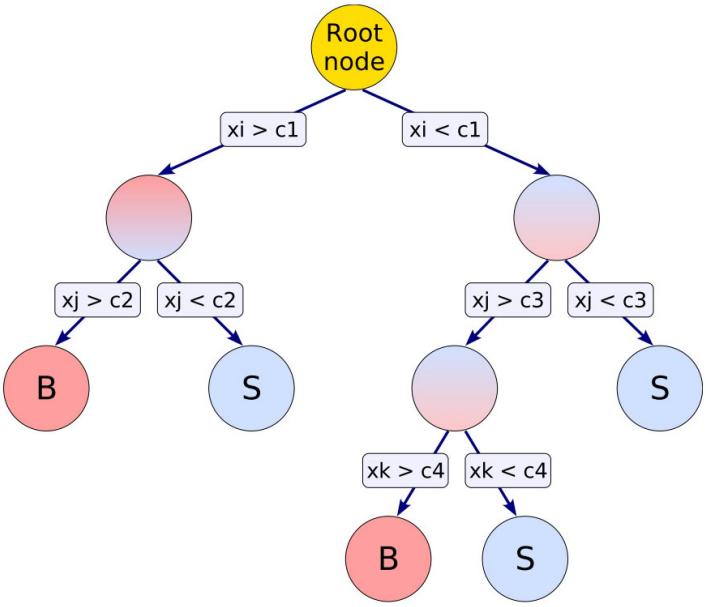



Рисунок 2 — Схематичный вид дерева решений

## 1.3 Бустинг

Недостатком деревьев решений является их чувствительность к флуктуациям в исходных данных. Например, из-за флуктуации одной переменной в тренировочном наборе данных может сильно повлиять на структуру итогового дерева решений.

Этой проблемы можно избежать, прибегнув к бустингу [4]. Суть этого алгоритма заключается в создании леса деревьев решений. При последовательном создании каждого дерева веса событий тренировочного образца изменяются таким образом, чтобы максимизировать влияние на построение дерева тех переменных, которые были неправильно классифицированы на предыдущих шагах. При этом каждому дереву присваивается вес, который отражает его эффективность в разделении событий.

При применении классификатора к набору данных, события поступают на вход каждому дереву решений, его отклик равен 1, если событие сигнальное и -1, если фоновое. Откликом классификатора является взвешенная сумма откликов всех деревьев в лесу.

## 1.4 Оценка результатов работы алгоритма

Для оценки эффективности работы алгоритма используются значения значимости при применении ограничений к отклику классификатора, определяемой выражением (1.1), а также площадь под ROC-кривой, которая является функцией зависимости эффективности отбора сигнала, определяемой выражением (1.2) и фонового отклонения, определяемого выражением (1.3) как функций от значения ограничения по отклику. Таким образом, чем больше площадь под ROC-кривой, тем эффективнее алгоритм отделяет сигнальные события от фоновых. Эффективность сигнала определяется как доля сигнальных событий, которая остаётся после применения классификатора. Отклонение фона – это доля фоновых событий, исключаемых из исходного набора.

$$\sigma = \frac{S}{\sqrt{S+B}} \tag{1.1}$$

$$\varepsilon = \frac{S}{S_{\text{init}}} \tag{1.2}$$

$$\kappa = 1 - \frac{B}{B_{\text{init}}} \tag{1.3}$$

где $S$ – число сигнальных событий, $B$ – число фоновых событий, $S_{\text{init}}$ и $B_{\text{init}}$ – число сигнальных и фоновых событий в исходном наборе соответственно.

Результат работы классификатора сравнивается с результатом классического метода с фиксированными отборами по переменным. Этот метод заключается в последовательном применении ограничений по выбранным переменным и определением порога при котором будет обеспечен максимум значимости. Для сравнения с методом BDT может использоваться максимальная значимость итоговых отборов, а также на график ROC кривой может быть нанесена точка, соответствующая значениям фиксированных отборов. Положение этой точке ниже кривой может свидетельствовать о меньшей эффективности классического метода.

# 2 Используемые данные

## 2.1 Устройство детектора

ATLAS [5] – это многоцелевой $4\pi$-детектор, один из четырёх крупнейших детекторов на Большом адронном коллайдере, расположенном в Европейской организации по ядерным исследованиям CERN, Женева, Швейцария. Он состоит из внутреннего трекового детектора, электромагнитного и адронного калориметров, магнитной системы а также мюонной системы. Устройство детектора ATLAS изображено на рисунке 3. Трековый детектор предназначен для определения треков заряженных частиц для измерения их импульса. калориметры необходимы для измерения энерговыделения частиц, мюонная система используется для определения импульса и направления пролёта мюонов. Магнитная система необходима для искривления траекторий заряженных частиц для определения их импульса.

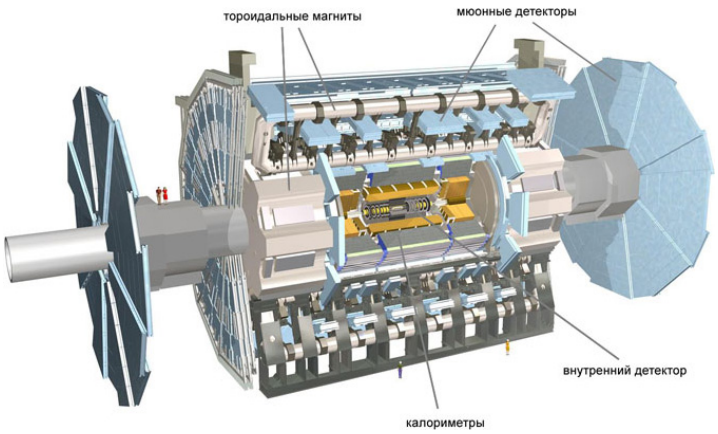



Рисунок 3 — Устройство детектора ATLAS

Триггерная система детектора состоит из нескольких уровней. Она снижает частоту событий с десятков мегагерц до сотен герц, отбирая события, представляющие интерес для анализа.

Для описания направления вылета частиц, используется цилиндрическая система координат. Азимутальный угол $\phi$ отсчитывается в плоскости, перпендикулярной оси детектора, полярный $\theta$ – от положительного

направления оси $z$, направленной вдоль оси детектора. Обычно вместо угла $\theta$ используется величина псевдобыстроты (2.1), дающая более равномерное распределение частиц, рождённых при столкновении.

$$\eta = -\ln \operatorname{tg} \frac{\theta}{2} \tag{2.1}$$

## 2.2 Исходные данные

Работа проводится с данными, полученными методом Монте-Карло моделирования протон-протонного столкновения в детекторе ATLAS на Большом адронном коллайдере с энергией в системе ценра масс 13 TeV и интегральной светимости 139 $\text{fb}^{-1}$. К исходным данным применены ограничения для отбора кандидатов на процессы с конечным состоянием, содержащим Z-бозон, фотон и две адронные струи (Z-бозон распадается на нейтрино и антинейтрино, которые дают недостающий поперечный импульс). Ограничения перечислены в таблице 1. Ограничения накладываемые на $E_T^{miss}$ и $E_T^{\gamma}$ являются триггерными. Условия на число фотонов, струй соответствует конечному состоянию процесса. Лептонное вето отсеивает процессы с лептонами в конечном состоянии. Угловые ограничения оптимизированы таким образом, чтобы максимально подавлять прочие фоны.

Таблица 1 — Критерии отбора событий

| Переменная | Ограничение |
|---|---|
| $E_T^{miss}$ | >120 GeV |
| $E_T^{\gamma}$ | >150 GeV |
| Число фотонов | $N_\gamma = 1$ |
| Число струй | $N_{jets} \geq 2$ |
| Число лептонов | $N_e = 0, N_\mu = 0$ |
| $\lvert\Delta\phi(\gamma, \vec{p}_T^{\,miss})\rvert$ | $> 0.4$ |
| $\lvert\Delta\phi(j_1, \vec{p}_T^{\,miss})\rvert$ | $> 0.3$ |
| $\lvert\Delta\phi(j_2, \vec{p}_T^{\,miss})\rvert$ | $> 0.3$ |

# 3 Процесс работы и результаты

## 3.1 Распределения по кинематическим переменным

Были построены распределения по переменным, показавшим наибольшую эффективность в разделении сигнала и фона. К ним относятся инвариантаная масса двух струй $m_{jj}$, разность быстрот двух струй $\Delta Y(j_1, j_2)$, недостающий поперечный импульс $E_T^{\text{miss}}$, баланс поперечных импульсов $p_T$ − balance, определяемый выражением (3.1), псевдобыстрота второй по значению поперечного импульса струи $\eta(j_2)$, поперечный импульс лидирующей струи $p_T(j_1)$, псевдобыстрота фотона $\eta(\gamma)$, сокращённый баланс поперечных импульсов $p_T$ − balance(reduced), определяемый выражением (3.2), число адронных струй $N_{\text{jets}}$, $\sin|\frac{\Delta\varphi(j_1,j_2)}{2}|$ и разность псевдобыстрот между струёй и фотоном $\Delta Y(j_1, \gamma)$. Распределения представлены на рисунках 4-5. В качестве фоновых событий использовалась сумма всех фонов, рассматриваемых в этом анализе.

$$p_T - \text{balance} = \frac{|\vec{p}_T^{\,miss} + \vec{p}_T^{\,\gamma} + \vec{p}_T^{\,j_1} + \vec{p}_T^{\,j_2}|}{E_T^{miss} + E_T^{\,\gamma} + p_T^{\,j_1} + p_T^{\,j_2}} \tag{3.1}$$

$$p_T - \text{balance(reduced)} = \frac{|\vec{p}_T^{\,\gamma} + \vec{p}_T^{\,j_1} + \vec{p}_T^{\,j_2}|}{E_T^{\gamma} + p_T^{\,j_1} + p_T^{\,j_2}} \tag{3.2}$$

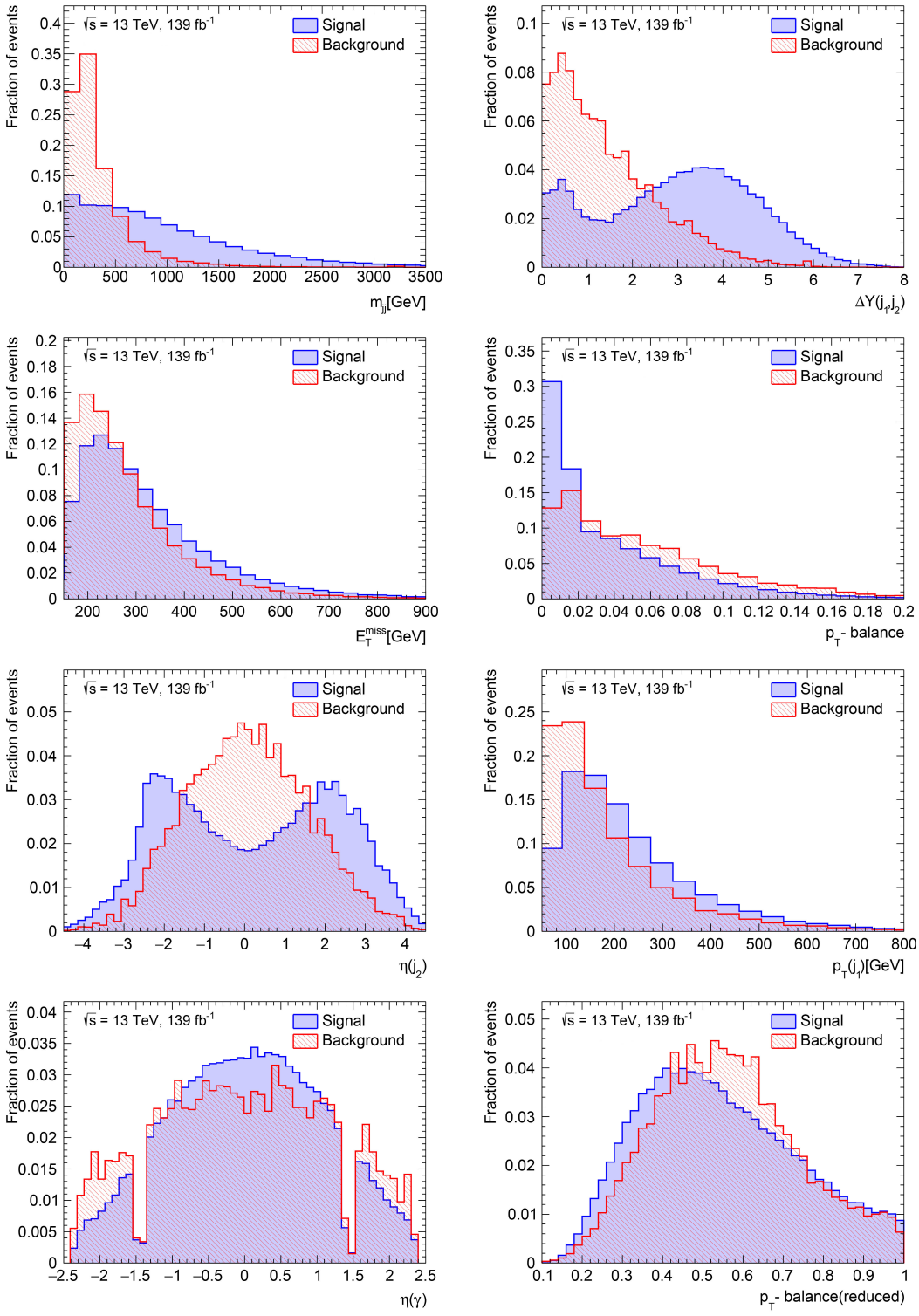



Рисунок 4 — Распределения по рассматриваемым в работе переменным, нормированным на полное число событий, для сигнала и фона

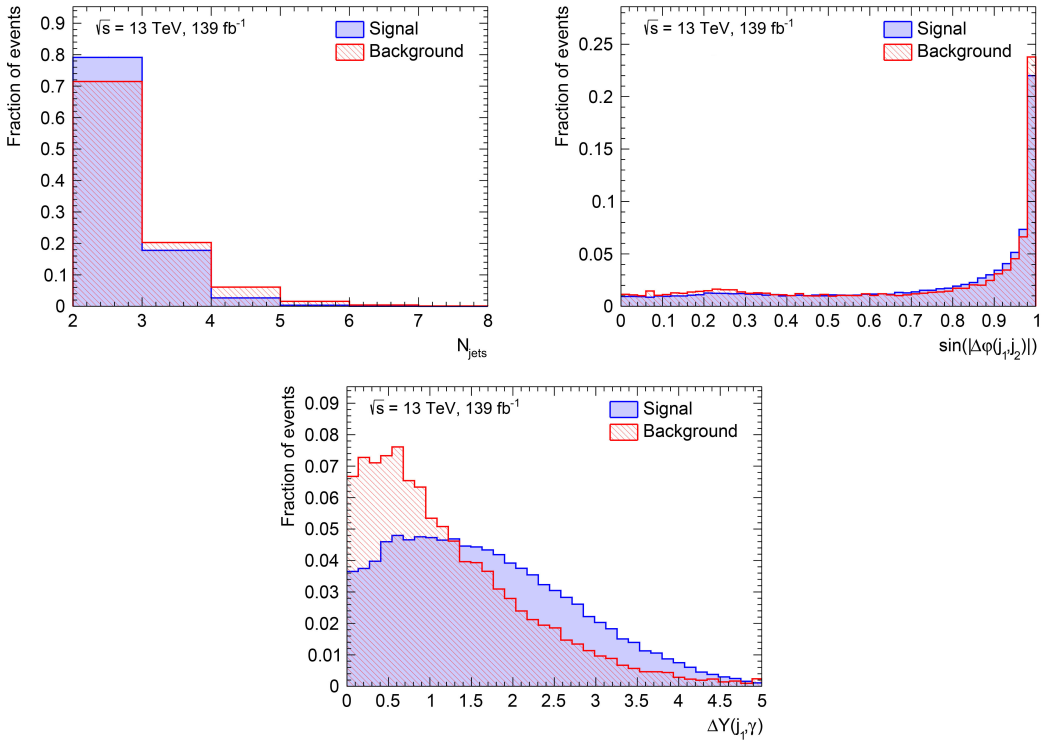


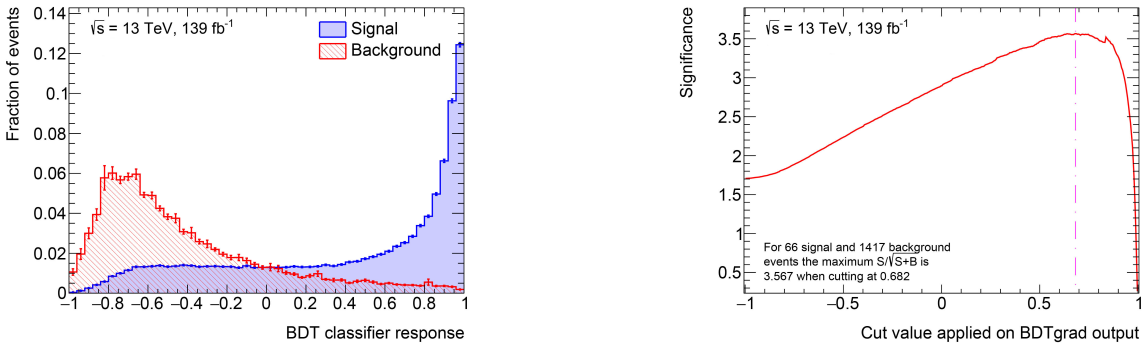

Рисунок 5 — Распределения по рассматриваемым в работе переменным, нормированным на полное число событий, для сигнала и фона

## 3.2 Создание и применение классификатора

В качестве модели для обучения использовался BDT классификатор, настройки которого перечислены в таблице 2. Он создавался при помощи библиотеки машинного обучения TMVA [3], входящей в программный пакет ROOT [6]. После применения классификатора к исходному набору данных было построено распределение отклика, график значимости, которая определяется выражением (1.1) как функции от ограничения по отклику классификатора и определено положение его максимума. Они изображены на рисунках 6а и 6б соответственно. Также была построена ROC-кривая, и определена площадь под ней. Её график изображен на рисунке 7. Результаты представлены в таблице 3.

Таблица 2 — Настройки классификатора

| | |
|---|---|
| Количество деревьев | 850 |
| Максимальная глубина | 3 |
| Тренировочные данные | 50% |




(а) (б)

Рисунок 6 — (а) Распределение отклика классификатора и (б) график значимости как функции от ограничения по отклику классификатора

## 3.3 Фиксированные отборы по переменным

Для сравнения эффективности применения методов машинного обучения были проведены фиксированные отборы по переменным центральности фотона $\zeta(\gamma)$, определяемой выражением (3.4), инвариантной массы двух струй $m_{jj}$, баланса поперечных импульсов $p_T-$balance и разница псевдобыстрот струй $\Delta Y(j_1, i_2)$, которые хорошо показали себя в предыдущем исследовании [1]. Была произведена оптимизация значимости для каждой комбинации ограничений. Самая большая значимость была получена для следующего порядка применения ограничений

$$\zeta(\gamma) \to m_{jj} \to p_T - balance \to \Delta Y(j_1, i_2) \tag{3.3}$$

$$\zeta(\gamma) = |\frac{\eta_\gamma - \frac{\eta_{j_1}+\eta_{j_2}}{2}}{\eta_{j_1} - \eta_{j_2}}| \tag{3.4}$$

где $\eta_\gamma$ – псевдобыстрота фотона, $\eta_{j_1}$, $\eta_{j_2}$ – псевдобыстроты струй.

Значения ограничений приведены в таблице 4. Графики значимости, эффективности сигнала и фона для фиксированных отборах приведены на рисунке 8. Максимум значимости представлен в таблице 3. Полученное значение меньше значимости, полученной при применении классификатора на величину более одного стандартного отклонения. Для этих отборов также была получена точка на графике ROC-кривой. Иллюстрация взаимного расположения этой точки и ROC-кривой изображено на рисунке 7. Точка

оказалась под кривой, что также может подтверждать большую эффективность классификатора. Количества событий после применения классификатора и одномерных фиксированных отборов представлены в таблице 5.

Таблица 3 — Результаты применения классификатора и одномерных фиксированных отборов

| | |
|---|---|
| $\mathrm{Max}(\sigma_{\mathrm{class}})$ | $3.57 \pm 0.06$ |
| $\mathrm{Max}(\sigma_{\mathrm{fix}})$ | $3.23 \pm 0.06$ |
| ROC-area | 0.86 |

Таблица 4 — Порядок ограничений, обеспечивающий максимальную значимость

| Переменная | Ограничение |
|---|---|
| $\zeta(\gamma)$ | $< 0.455$ |
| $m_{jj}$ | $> 697$ GeV |
| $p_T - balance$ | $< 0.064$ |
| $\Delta Y(j_1, i_2)$ | $> 2.227$ |

Таблица 5 — Количества событий до и после применения отборов

| | Сигнал | Фон |
|---|---|---|
| До применения отборов | 66 | 1417 |
| Классификатор | 30 | 42 |
| Фиксированные отборы | 26 | 40 |

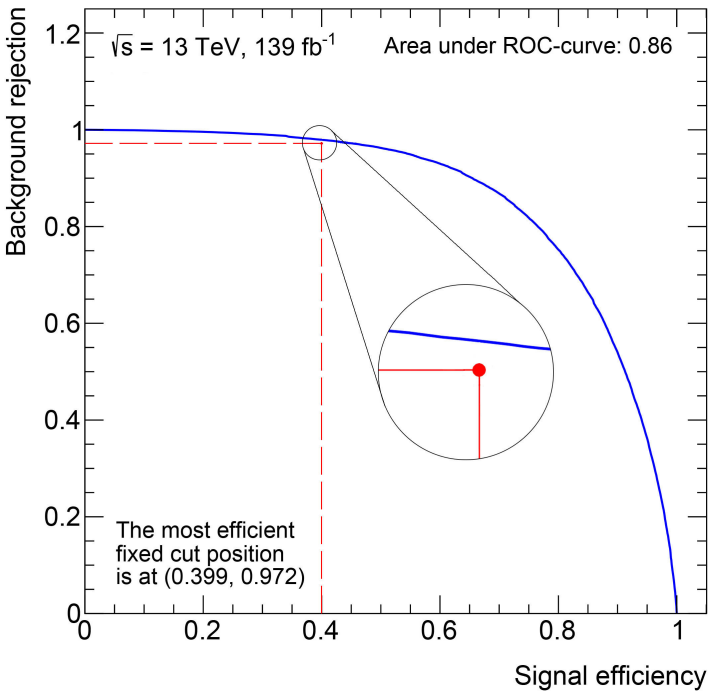



Рисунок 7 — ROC-кривая и точка, соответствующая фиксированному отбору, обеспечивающему максимальную значимость

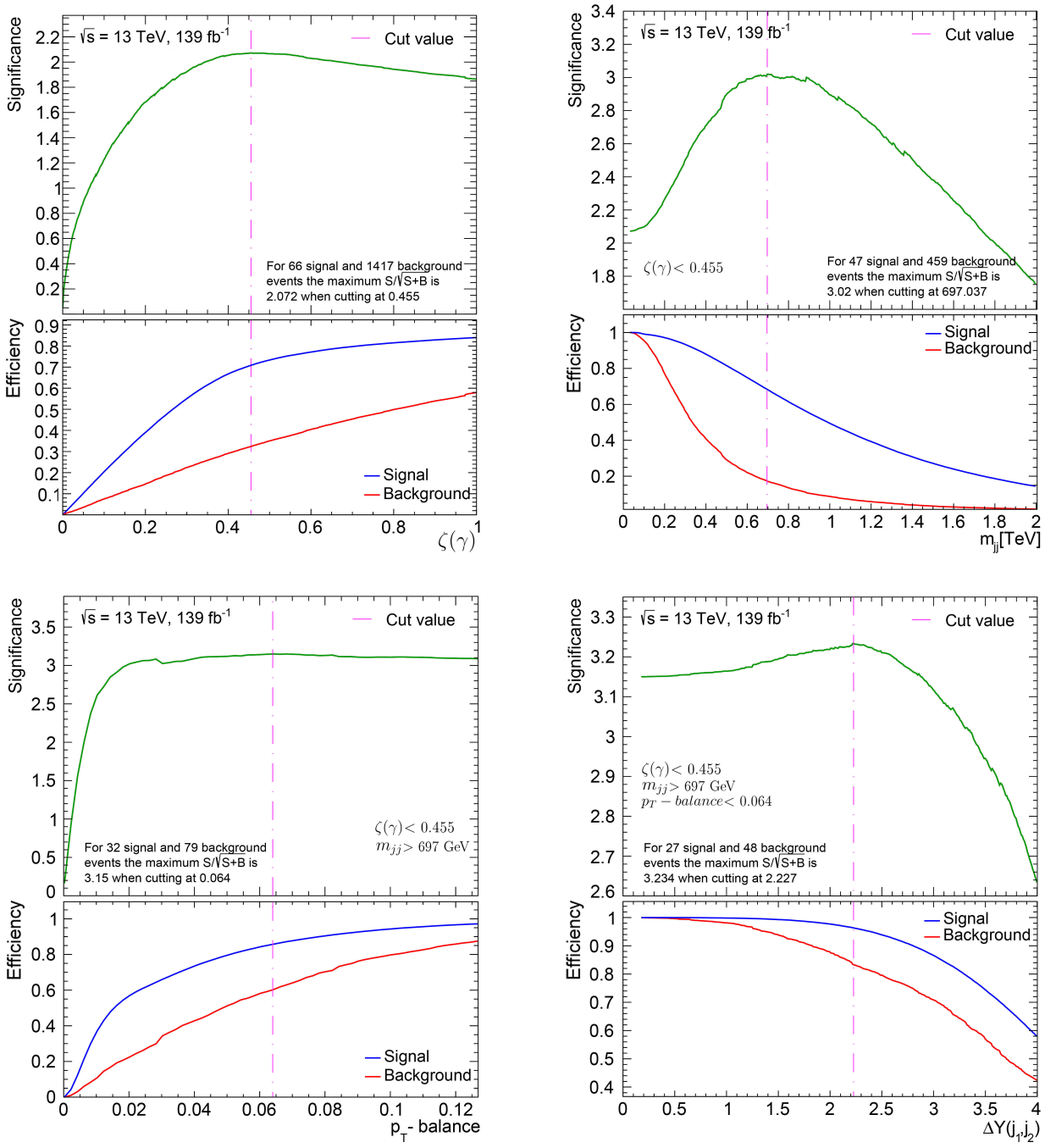



Рисунок 8 — Зависимости значимости, эффективности сигнала и фона от ограничений по переменным при фиксированных отборах для порядка отборов, обеспечивающего максимальную значимость

# Заключение

В процессе работы было произведено сравнение эффективности разделения сигнальных и фоновых событий при применении метода BDT и фиксированных отборов по наиболее значимым переменным. Результат показал явный прирост значимости при использовании классификатора по сравнению с классическими отборами по переменным.

В дальнейшем планируется проверка эффективности других алгоритмов машинного обучения, их настройка и определение наилучших переменных для использования их в обучении. Также для дальнейшего использования классификатора необходимо определить насколько Монте-Карло моделирование согласуется с реальными данными от столкновений.

# Список используемых источников


1. *Aaboud M.* [et al.]. Studies of Z production in association with a high-mass dijet system in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector // Journal of High Energy Physics. — 2017. — July. — Vol. 2017, no. 7. — ISSN 1029-8479. — URL: http://dx.doi.org/10.1007/JHEP07(2017)107.

2. *Khachatryan V.* [et al.]. Measurement of the cross section for electroweak production of Z in association with two jets and constraints on anomalous quartic gauge couplings in proton–proton collisions at s=8 TeV // Physics Letters B. — 2017. — Vol. 770. — P. 380–402. — ISSN 0370-2693. — URL: http://www.sciencedirect.com/science/article/pii/S0370269317303453.

3. *Hoecker A.* [et al.]. TMVA - Toolkit for Multivariate Data Analysis. — 2007. — arXiv: physics/0703039 [physics.data-an].

4. *Friedman J. H.* Greedy function approximation: A gradient boosting machine // Ann. Stat. — 2001. — Vol. 29, no. 5. — P. 1189–1232. — ISSN 0090-5364; 2168-8966/e.

5. *Collaboration T. A.* [et al.]. The ATLAS Experiment at the CERN Large Hadron Collider // Journal of Instrumentation. — 2008. — Aug. — Vol. 3, no. 08. — S08003–S08003. — URL: https://doi.org/10.1088/1748-0221/3/08/s08003.

6. *Brun R., Rademakers F.* ROOT — An object oriented data analysis framework // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. — 1997. — Vol. 389, no. 1. — P. 81–86. — ISSN 0168-9002. — URL: http://www.sciencedirect.com/science/article/pii/S016890029700048X ; New Computing Techniques in Physics Research V.