МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ
УНЕВЕРСИТЕТ «МИФИ»

(НИЯУ МИФИ)

УДК 539

ОТЧЕТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

ОПТИМИЗАЦИЯ ОТБОРОВ В $ZZ \to ll\nu\nu$ АНАЛИЗЕ

Руководитель НИРС,
к.ф-м.н. _____ Солдатов Е.Ю.
подпись

Студент гр. М20-115 _____ Зубов Д.В.
подпись

Консультант _____ Пятиизбянцева Д.Н.
подпись

Консультант _____ Петухов А.М.
подпись

Москва 2022

Содержание

Введение

Стандартная модель (СМ) физики элементарных частиц объясняет большинство явлений и процессов в физике высоких энергий, а ее предсказания подтверждались во множестве экспериментах. Однако, Стандартная модель считается неполной, поскольку она не отвечает на многие фундаментальные вопросы. В связи с чем предполагается, что СМ является частью более универсальной теории и обнаружение отклонений от предсказаний СМ может подтвердить или отбросить новые теории.

Рождение пар векторных бозонов тесно связано с неабелевой природой электрослабой теории и спонтанным нарушением калибровочной симметрии. Кроме того, предсказывается широкий спектр новых явлений за пределами Стандартной модели (СМ) физики частиц, связанный с рождением двубозонной пары. Изучение процессов рождения векторных бозонов является краеугольным камнем электрослабой теории и возможных сценариев физики за пределами СМ и составляет существенную часть физической программы Большого адронного коллайдера (БАК).

Среди всех двубозонных процессов рождение пары Z-бозонов имеет наименьшее сечение, но, тем не менее, процесс вполне перспективен для измерения параметров СМ и поиска «новой» физики благодаря хорошему соотношению сигнал/фон в канале распада на четыре заряженных лептона. Соотношение сигнал/фон несколько хуже в канале распада на пару заряженных лептонов и пару нейтрино, но вероятность таких распадов выше[1].

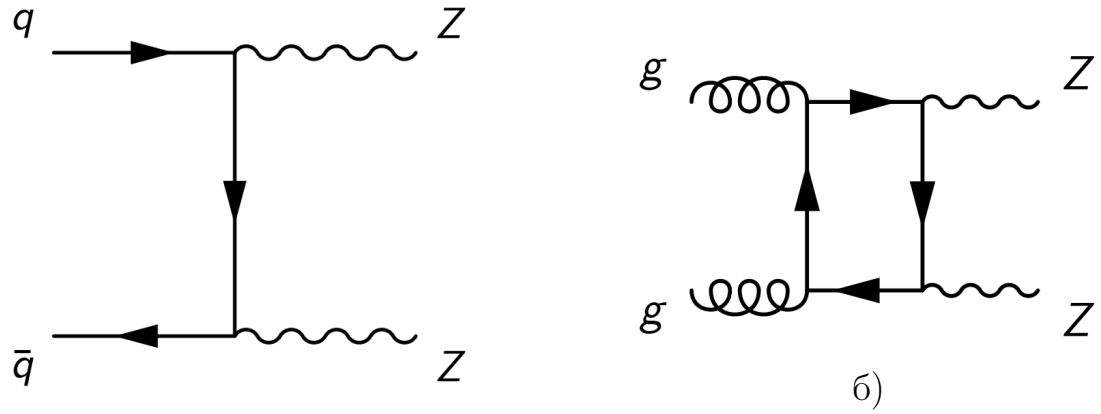


Рис. 1: Диаграммы процессов КХД рождения двух Z бозоно

В работе показана оптимизация отбора событий с использованием алгоритмов машинного обучения для инклюзивного рождения пары Z-бозонов и последующего распада на два заряженных лептона и два нейтрино. В результате оптимизации может быть достигнута максимальная значимость отбора событий, что позволит измерить сечение рождения пары Z бозонов с большей точностью.

1 Экспериментальная установка ATLAS

Эксперимент ATLAS[2] (ATLAS — A Toroidal LHC ApparatuS) многоцелевой детектор, покрывающий почти полный телесный угол. В эксперименте ATLAS используется прямоугольная система координат. Ось z направлена по оси пучка, x - к центру кольца, y - вверх, ϕ - азимутальный угол в плоскости xOy, перпендикулярной пучку, отсчитывается от оси x, θ - полярный угол, отсчитывается от оси Z . В основном при работе используется величина, зависящая от полярного угла $\eta = -\ln\left(\mathrm{tg}\,\frac{\theta}{2}\right)$, называемая псевдобыстротой, т.к. она аддитивна относительно преобразований Лоренца.

Эксперимент ATLAS включает в себя внутренний детектор (ВД), систему калориметров, мюонный спектрометр (МС), магнитную и тригерную системы(Рис. 1).

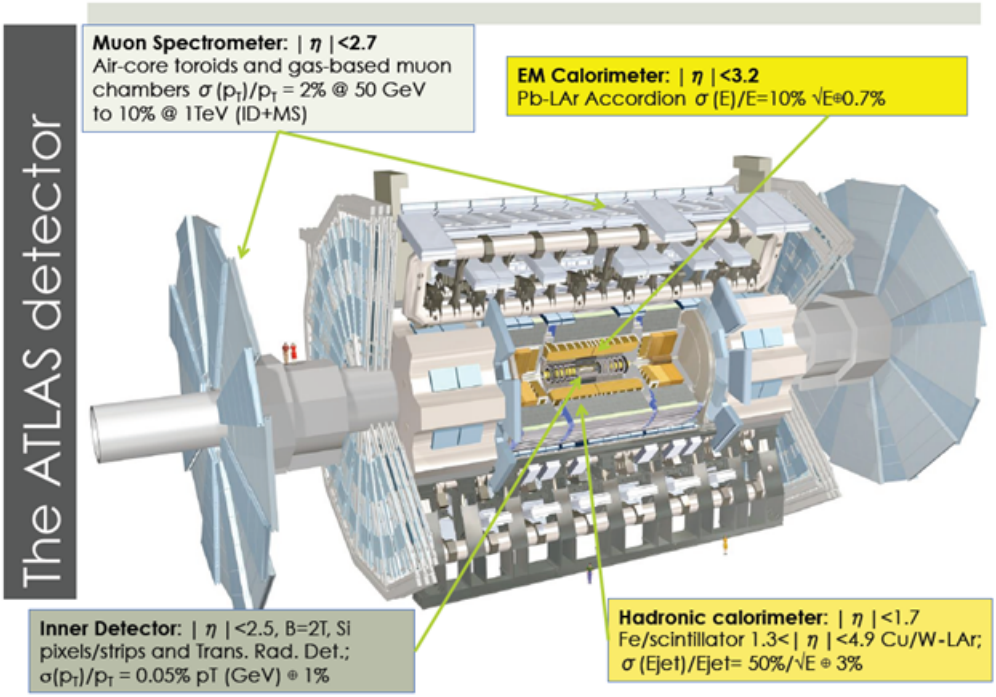


Рис. 2: Детектор ATLAS

Основная задача внутреннего детектора - восстановление треков заряженных частиц. ВД состоит из трех подсистем : пиксельного (Pixel) и силиконового (SCT) микростриповых детекторов, а также трекера переходного излучения (TRT). Пиксельный детектор состоит из трех цилиндрических слоев и трех торцевых пластин и в общем содержит 80 миллионов пикселей. Ближайший к пучку слой называют B-слоем. Он находится на расстоянии 3.3 см от пучка и имеет важную роль в востановлении треков. SCT детектор включает 4 цилиндреческих слоя и 9 дисков в каждом торце, состоящих из силиконовых микрострипов. Силиконовый и пиксельный детекторы покрывают область псевдобыстроты $|\eta < 2.5|$. TRT, состоящий из тонких трубок,

напоненных смесью Xe и Ar, покрывает область псевдобыстрот $|\eta < 2.0|$.

Калориметрическая система состоит из электромагнитной (ЭМК) и адронной составляющей. ЭМК играет решающую роль в идентификации электронов и фотонов. Он состоит из слоев свинца и жидкого аргона и имеет геометрию аккордеона. ЭМК делится на центральную часть, покрывающую область псевдобыстрот $|\eta| < 1.475$, и две торцевые части (каждая из которых состоит из двух коаксиальных колёс), покрывающие область псевдобыстрот$1.375 < |\eta| < 3.2$. В области псевдобыстрот $1.37 < |\eta| < 1.52$ находится технический зазор, в котором измерения не проводятся. Адронный калориметр состоит из 3-х различных систем: Tile-калориметр, торцевой LAr-калориметр и передний LAr-калориметр. Tile-калориметр размещается снаружи корпуса ЭМ-калориметра. Он состоит из органических сцинтилляторов и позволяет регистрировать энергии адронов в области с псевдобыстротой $|\eta < 1.7|$. Торцевой LAr-калориметр, рабочим веществом которого является жидкий аргон, расположен за торцевым ЭМ-калориметром. Он перекрывает область псевдобыстрот $1.5 < |\eta| < 3.2$. Передний LAr-калориметр, также основанный на жидком аргоне, создает однородность калориметрии и поглощает фон перед мюонными камерами. Его область псевдобыстрот: $3.1 < |\eta| < 4.9$.

МС востанавливает импульс и треки пролетающих мюонов с максимально возможным разрешением. Состоит из четырёх подсистем, использующих разные технологии: Мониторируемые Дрейфовые Трубки, Катодные Стриповые Камеры, Резистивные Плоские Камеры и Тонко-Зазорные Камеры. Эти подсистемы погружены в магнитное поле, генерируемое тремя тороидами: один центральный покрывает диапазон по псевдобыстроте $|\eta < 1.5|$ обеспечивая поле в 0.5 Тл и ещё два, расположенные в области большей псевдобыстроты $|\eta > 1.5|$ генерируют поле в 1 Тл.

Для предварительного отбора «интересных» столкновений используется система триггеров. В результате её,при номинальной частоте столкновений 40 МГц, интересные события поступают со средней частотой 200 Гц.

Для измерения импульсов создана специальная система магнитов, создающая электромагнитное поле, которое искривляет траектории заряженных частиц. Она состоит из 4-х сверхпроводящих магнитов: соленоида и трёх тороидов. Подразделяется на 2 основных составляющих – внутреннюю (соленоид) и внешнюю (тороидальные магниты).

2 Оптимизация отбора событий

2.1 Алгоритм BDTG

Оптимизация отбора событий происходила с использования алгоритма BDTG реализованного в пакете TMVA [4]. BDTG - это «Композиция деревьев решений» (Boosted Decision Trees) [5] использующая градиентный бустинг [6,7]. Принцип работы алгоритма состоит в поочерёдном применении ограничений по различным переменным, в ходе чего строится дерево решений. Отборы по переменным производятся 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 каком-либо 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.

Недостатком деревьев решений является их чувствительность к флуктуациям в исходных данных и склонность к перетренированности. Бустинг решает эту проблему . 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. Отклик классификатора – непрерывная величина, лежащая в пределах [-1;1] и являющаяся взвешенной суммой откликов всех деревьев в лесу. Распределение по отклику можно использовать для разделения сигнальных и фоновых событий.

Для оценки эффективности разделения сигнала и фона классификатором использовались сигнальная значимость (1) и площадь под ROC-кривой, которая является функцией зависимости эффективности отбора сигнала (signal efficiency) (2) и фонового подавления (background rejection) (3) как функций от значения ограничения по отклику. Эффективность сигнала определяется как доля сигнальных событий, которая остаётся после применения классификатора. Подавление фона – это доля фоновых событий, исключаемых из исходного набора.

$$Z = \sqrt{2 \times [(S+B) \times \ln(1+(S/B)) - S]}, \tag{1}$$

где Z - сигнальная значимость, S - число сигнальных событий, B - число фоновых событий.

$$\varepsilon = \frac{S}{S_{init.}}, \tag{2}$$

$$\kappa = 1 - \frac{B}{B_{init.}}, \tag{3}$$

где $S_{init.}$ и $B_{init.}$ – число сигнальных и фоновых событий в исходном наборе соответственно.

2.2 Оптимизация отборов в процессе инклюзивного рождения пары Z-бозонов и последующего распада на пару заряженных лептонов и пару нейтрино.

Среди всех двубозонных процессов рождение пары Z-бозонов имеет наименьшее сечение, но хорошее соотношение сигнал/фон в канале распада на четыре заряженных лептона. Соотношение сигнал/фон несколько хуже в канале распада на пару заряженных лептонов и пару нейтрино,но вероятность таких распадов выше.

Сигнатурой этого процесса в детекторе ATLAS являются события, содержащие пару разноименно заряженных лептонов (e^+e^- или $\mu^+\mu^-$) и большой потерянный поперечный импульс, который соответствуют Z-бозону, распавшемуся на пару нейтрино. Схожую сигнатуру имеет ряд других фоновых процессов.

Оптимизация отборов проводилась на данных Монте-Карло симуляции работы детектора ATLAS в течении второго сеанса набора данных. Сигнальные и фоновые процессы описаны в таблице 1.

Кандидаты в сигнальные события должны удовлетворять следующим критериям:

- В событии два разноименно-заряженных лептона одного аромата (e+e- или mu+mu-), при этом, поперечный импульс первого больше 30 ГэВ, второго больше 20 ГэВ;
- Вето на третий заряженный лептон;
- 76 ГэВ $< M_{ll} <$ 106 ГэВ, где M_{ll} - инвариантная масса двух заряженных лептонов;
- $E_T^{miss} > 70$ ГэВ.

Сигнал	
QCD ZZ	КХД рождение двух Z-бозонов и последующий распад в $ll\nu\nu$
EWK ZZ	Электрослабое рождение двух Z-бозонов и последующий распад в $ll\nu\nu$
Фон	
Zj	рождение Z-бозона и струи, с распадом Z-бозона в пару заряженных лептонов и большим ложным потерянным поперечным импульсом
WZ	рождение пары бозонов Z и W, с распадом Z-бозона в пару заряженных лептонов и лептонным распадом W
tt	рождение пары топ-кварков и последующим распадом включающим конечное состояние $ll\nu\nu$ (не резонансное рождение $ll\nu\nu$)
WW	рождение пары W с распадом в $ll\nu\nu$ (не резонансное рождение $ll\nu\nu$)
Wt	рождение W и топ-кварка и распадом в конечное состояние, содержащее $ll\nu\nu$ (не резонансное рождение $ll\nu\nu$)
VVV	рождение трех векторных бозонов (V = W или Z)
Other (ttV, ttVV)	рождение пары топ-кварков и одного или двух векторных бозонов

Таблица 1: Сигнальные и основные фоновые процессы для процесса инклюзивного рождения ZZ и последующего распада в $ll\nu\nu$

2.3 Предотбор событий «жесткими» условиями

Перед непосредственным использованием алгоритма BDTG проводился предварительный отбор событий ограничениями на переменные, найденными в ходе cut based оптимизации переменных. В таблице 2 представлены отборы, полученные в ходе cut based оптимизации, значение сигнальной значимости, отношение числа сигнальных событий к фоновым, число фоновых и сигнальных событий, до и после оптимизации. В таблице 3 представлены числа сигнальных и фоновых процессов для каждого источника сигнала и фона до и после cut based оптимизации.

Переменная	До	После
E_T^{miss} значимость	—	>10
E_T^{miss}, ГэВ	—	—
ΔR_{ll}	—	<1.8
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	—	>2.3
Число b-струй	—	<1
E_T^{miss}/H_T	—	>0.5
Сигнальная значимость	7.43 ±0.03	44.7 ±0.4
Сигнал/Фон	0.007	1.43
Число сигнальных событий	7858 ±28	1959 ±15
Число фоновых событий	(1123 ±4)·10^3	1370 ±22

Таблица 2: Результаты cut based оптимизации инклюзивного рождения ZZ

	До	После
Сигнал		
QCD ZZ	7596 ±28	1946 ±15
EWK ZZ	262 ±2	13.0 ±0.4
Total signal	7858 ±28	1959 ±15
Фон		
Zj	962833 ±4057	181 ±20
WZ	11338 ±29	945 ±8
tt	123340 ±73	131 ±2
WW	5093 ±13	64.0 ±1.5
Wt	10251 ±41	41 ±3
VVV	41.8 ±0.3	7.88 ±0.10
Other	282 ±2	0.79 ±0.11
Total bkg.	$(1123 \pm 4) \cdot 10^3$	1370 ±22

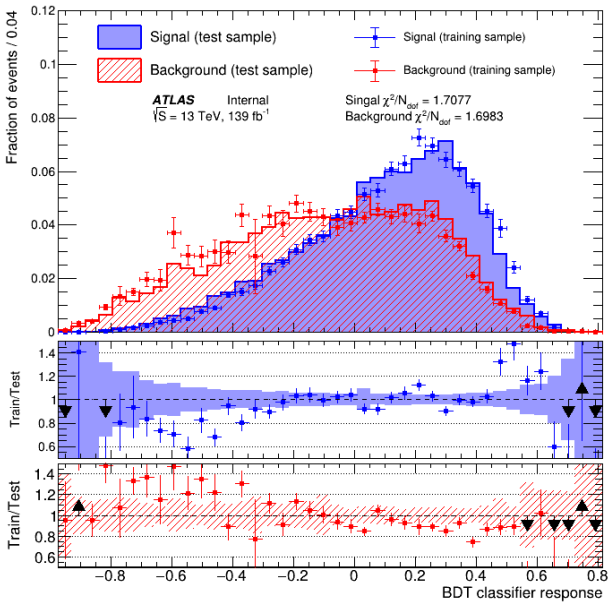
Таблица 3: Числа сигнальных и фоновых процессов для каждого источника сигнала и фона до и после cut based оптимизации.

Список переменных, их распределения и корреляционные матрицы представлены в приложении. На рисунке 4 показаны результаты работы классификатора. Из распределения по отклику классификатора и ROC-кривой видно низкая эффективность работы алгоритма на выделенном наборе данных. Полученный результат объясняется использованием «жестких» предварительных отборов, которые также являются результатом оптимизации.

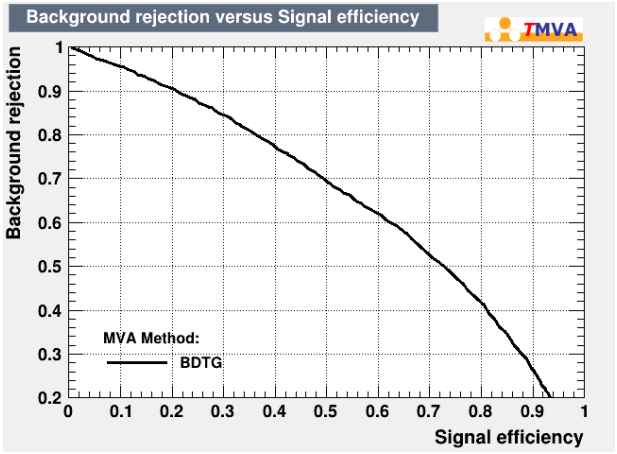
Тем не менее осуществлялись попытка улучшить работу классификатора удаляя скоррелированные переменные и фоны, вносящие большой вклад в статистическую ошибку числа событий. К положительным результатам эти действия не привели, подробное сравнение классификатора с разными опциями представлены в приложении.

2.4 Предотбор событий расслабленными условиями

В целях увеличения разделительной способности переменных и улучшения качества работы классификатора обучение производилось на событиях отобранных с помощью расслабленных условий на переменные. Эти отборы были получены с помощью метода cut based оптимизации поиском максимума сигнальной значимости при условии, что число сигнальных событий больше 4000. В таблице 4 представлены «жесткие» и расслабленные отборы, значение сигнальной значимости, отношение числа сигнальных событий к фоновым, число фоновых и сигнальных событий. В таблице 5 представлены числа сигнальных и фоновых процессов для каждого источника сигнала и фона для «жесктого» и расслабленного вариантов преселекции.



а)



б)

Рис. 3: Распределение сигнала и фона по переменной функции отклика классификатора слева и ROC-кривая справа при жесткой преселекции.

Переменная	Расслабленный отбор	Жесткий отбор
E_T^{miss} значимость	7	>10
ΔR_{ll}	<2.2	<1.8
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>2.3	>2.3
Число b-струй	<1	<1
E_T^{miss}/H_T	>0	>0.5

Таблица 4: Условия расслабленного и жесткого отбора событий

Процесс	Расслабленный отбор	Жесткий отбор
Сигнал		
QCD ZZ	4409 ±23	1946 ±15
EWK ZZ	57.8 ±0.9	13.0 ±0.4
Total signal	4467 ±23	1959 ±15
Фон		
Zj	12184 ±290	181 ±20
WZ	3116 ±15	945 ±8
tt	2829 ±11	131 ±2
WW	1352 ±7	64.0 ±1.5
Wt	729 ±10	41 ±3
VVV	1771 ±0.17	7.88 ±0.10
Other	4.46 ±0.26	0.79 ±0.11
Total bkg.	20439±291	1370 ±22

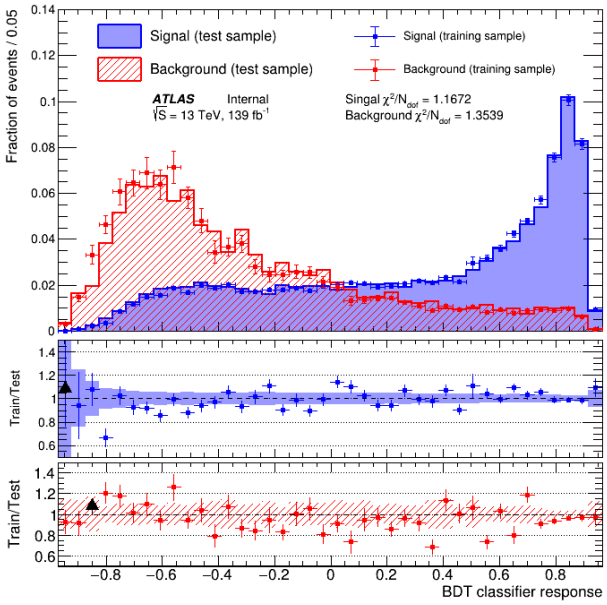
Таблица 5: Числа сигнальных и фоновых событий для каждого источника сигнала и фона при расслабленном и жестком отборе.

При обучении «расслабленного» классификатора использовались те же переменные, что и при обучении «жесткого». Распределения переменных при расслабленной преселекции представлено в приложении. На рисунке 4 показаны результаты работы классификатора. Из распределения отклика классификатора видна улучшенная разделительная способность классификатора. Сравнение распределения тренировочной и тестовой выборки критерием Пирсона показывает отсутствие переобучения.

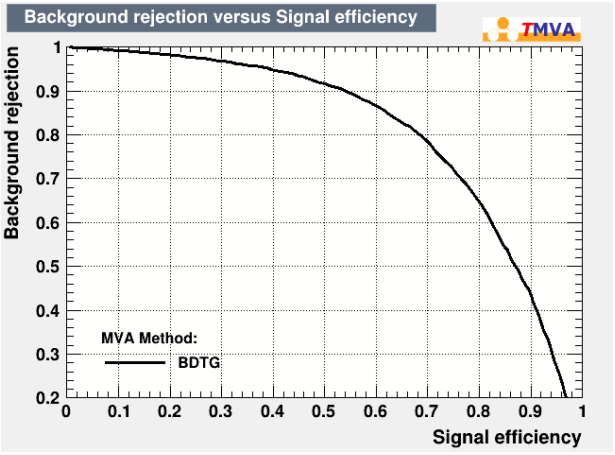
Аналогично производились сравнения классификаторов с удалением скоррелированных переменных. Из сравнения распределений отклика классификаторов видно, что при удалении скоррелированных переменных значение χ^2/NDF уменьшается, а значит согласие распределения тренировочной и тестовой выборки возрастает.

2.5 Оптимизация гиперпараметров классификатора.

Подбор оптимальных настроек (гиперпараметров) классификатора, таких как число деревьев, число бинов и глубина дерева, производился методом поиска по сетке. При этом в циклически классификатор обучался со всевозможными настройками классификатора и отбирался лучший, соответствующий максимуму площади под ROC-кривой. Сравнение классификатора с начальными параметрами и оптимальными показало незначительное увеличение сигнальной значимости и площади под ROC-кривой, не превышающее статистическую погрешность.



а)



б)

Рис. 4: Распределение сигнала и фона по переменной функции отклика классификатора слева и ROC-кривая справа при расслабленной преселекции.

3 Заключение

В ходе работы были рассмотрены различные варианты обучения классификатора. Было выяснено, что основное влияние на результат работы алгоритма оказывают данные, на которых он обучается, и используемые переменные. Путем расслабления начальных ограничений на переменные была увеличена разделяющая способность классификатора. В результате исключения скоррелированных переменных удалось улучшить стабильность отклика классификатора.

В результате был получен стабильный алгоритм, не склонный к переобучению. При этом, с помощью классификатора можно добиться увеличения сигнальной значимости с 44.1 ± 0.44 до 46 ± 0.3.

В дальнейшем планируется на основе полученного алгоритма точнее оценить число событий рождения пары Z бозонов и с большей прецизионностью измерить сечение данного процесса.

Список используемых источников

1. Measurement of ZZ production in the $llνν$ final state with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV / M. Aaboud // Journal of High Energy Physics.2019., No 10. ISSN 1029-8479.

2. ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 (2008) S08003.

3. ATLAS Computing : technical design report [Text]: Technical Design Report ATLAS (17) / ATLAS Collaboration. - Geneva : CERN, 2005. - 234 p.

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

5. L. Breiman, J. H. Friedman, R. A. Olshen, et al., Classification and regression trees, 1983.

6. Y. Freund and R. E. Schapire, A short introduction to boosting, in In Proceedings of theSixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1999, pp. 1401–1406

7. Friedman J. H. Greedy function approximation: A gradient boosting machine // Ann. Stat. — 2001. — Vol. 29, no. 5. — P. 1189–1232.

8. Sinervo P. K.Signal significance in particle physics // Conference onAdvanced Statistical Techniques in Particle Physics. 2002. с. 64-76. arXiv:hep-ex/0208005.

9. Observation of electroweak production of two jets and aZ -boson pair with the ATLAS detector at the LHC. 2020. arXiv:2004.10612 [hep-ex].

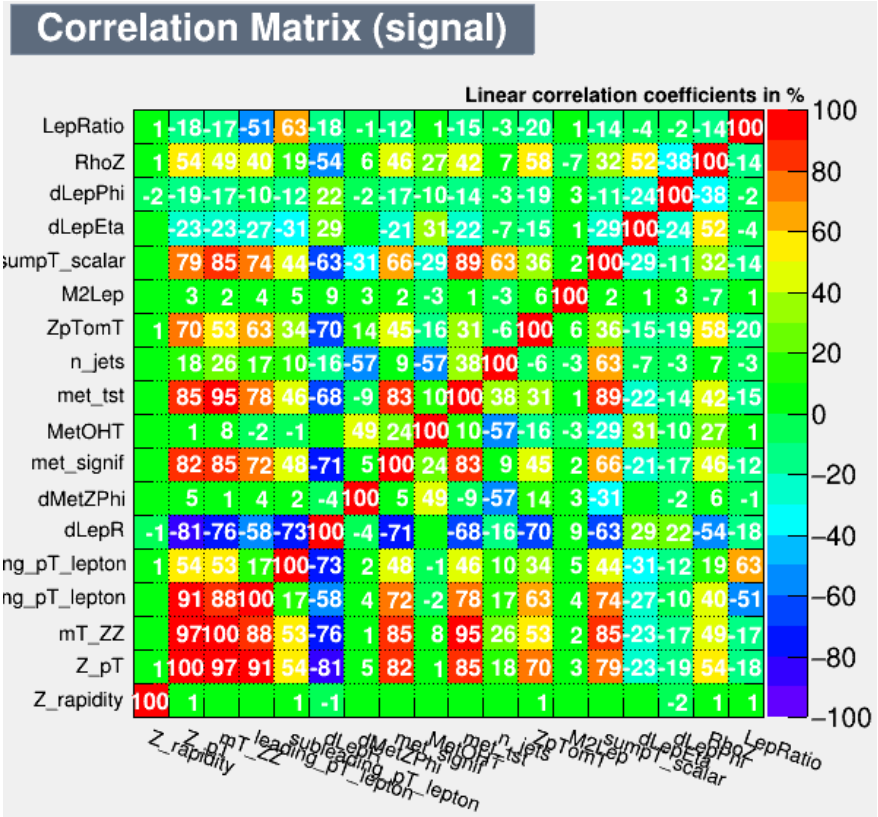
10. Collaboration A.Observation of electroweak production of two jets andaZ-boson pair with the ATLAS detector at the LHC. 2020. arXiv:2004.10612 [hep-ex].

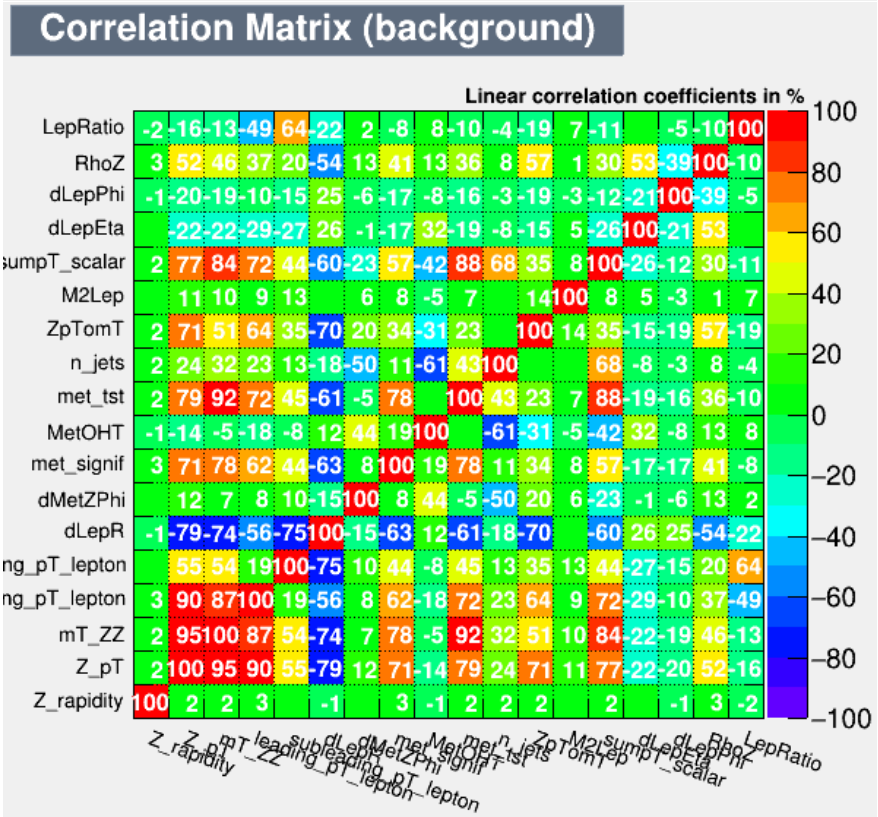
11. Rainwater D.,Szalapski R.,Zeppenfeld D.Probing color-singlet exchangeinZ+2-jet events at the CERN LHC // Physical Review D. 1996. т.54, No 11. с. 6680-6689. ISSN 1089-4918

1 Список используемых переменных

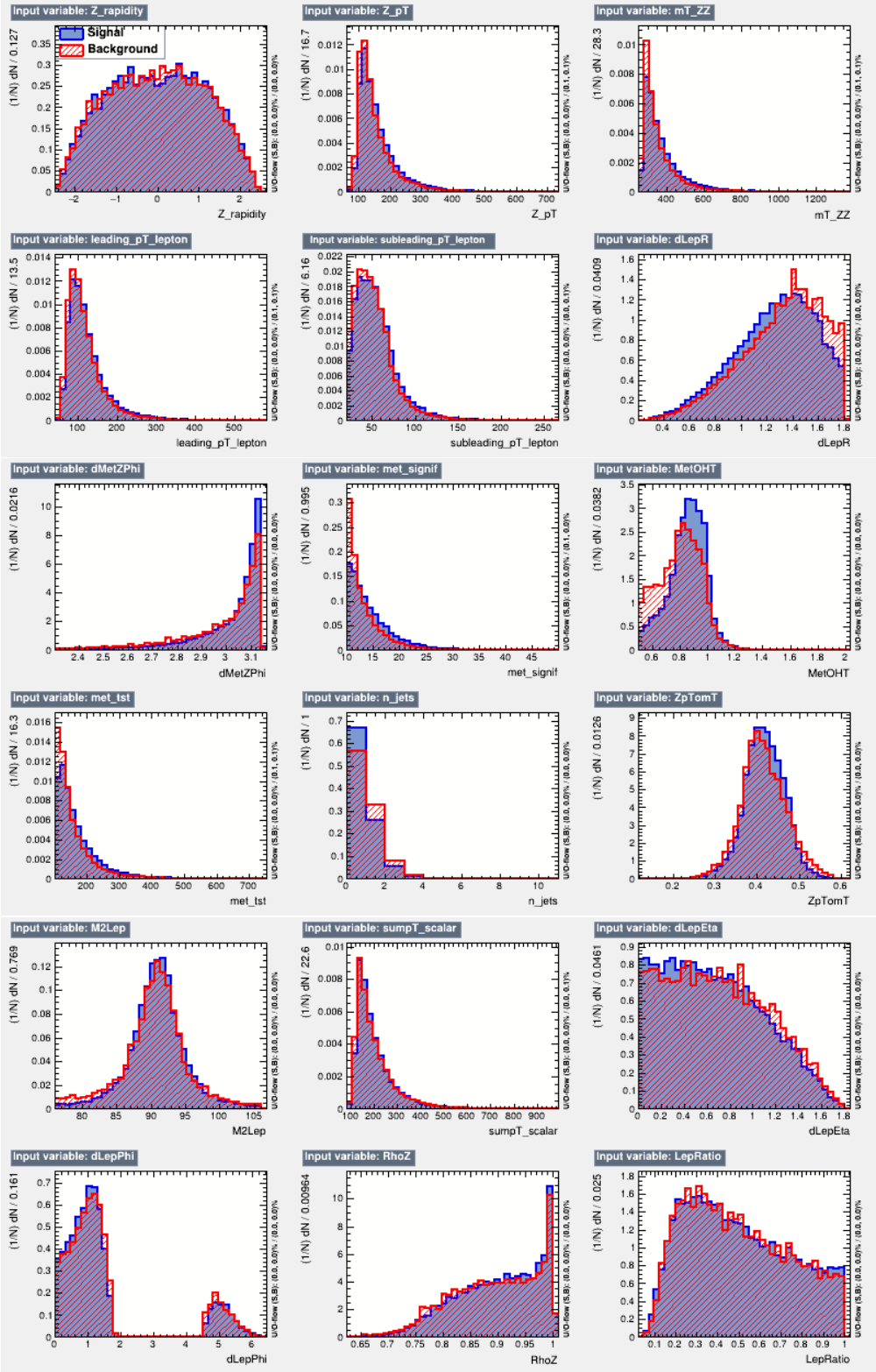
1. met_signif — E_T^{miss} значимость
2. met_tst — E_T^{miss} — потерянный попереный импульс
3. mT_ZZ — поперечная масса пары Z бозонов
4. dLepR — ΔR_{ll} - ΔR между двумя заряженными лептонами
5. leading_pT_lepton — поперечный импульс первого лептона
6. subleading_pT_lepton — поперечный импульс второго лептона
7. ZpTomT — отношение поперечного импульса Z бозона к его поперечной массе
8. MetOHT — отношение E_T^{miss}/H_T, где H_T скалярная сумма поперечных импульсов отобранных струй и заряженных лептонов
9. RhoZ — отношение поперечного импульса Z бозона к скалярной сумме поперечных импульсов лептонов
10. sumpT_scalar — скалярная сумма поперечных импульсов всех жестких объектов в событии
11. dLepPhi — $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$ - $\Delta\phi$ между лептонами
12. dMetZPhi — $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$ - $\Delta\phi$ между Z-бозонами
13. LepRatio — отношение поперечных импульсов лептонов
14. dLepEta — $\Delta\eta$ между двумя лептонами
15. M2Lep — инвариантная масса пары лептонов
16. Z_rapidity — быстрота Z бозона
17. n_jets — число струй в событии

2 Корреляционные матрицы переменных для сигнала и фона





3 Распределения используемых переменных. Жесткая преселекция.



4 Распределения используемых переменных. Расслабленная преселекция. Скоррелированные переменные исключены.