AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
Wydział Fizyki i Informatyki Stosowanej

Praca magisterska

Bartłomiej Pitala

Kierunek studiów**: Informatyka Stosowana**

Walidacja fizyczna powielonych przypadków rozpadu $B_s^0 \to D_s^* K^*$

Opiekun: dr hab. inż. **Tomasz Szumlak**

Kraków, listopad 2020 r.

Oswiadczenie studenta

Uprzedzony(-a) o odpowiedzialności karnej na podstawie art. 115 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz. U. z 2018 r. poz. 1191 z późn. zm.): „Kto przywłaszcza sobie autorstwo albo wprowadza w błąd co do autorstwa całości lub części cudzego utworu albo artystycznego wykonania, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 3. Tej samej karze podlega, kto rozpowszechnia bez podania nazwiska lub pseudonimu twórcy cudzy utwór w wersji orygi-nalnej albo w postaci opracowania, artystyczne wykonanie albo publicznie zniekształca taki utwór, artystyczne wykonanie, fonogram, wideogram lub nadanie.”, a także uprzedzony(-a) o odpowiedzialności dyscyplinarnej na podstawie art. 307 ust. 1 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2018 r. poz. 1668 z późn. zm.) „Student podlega odpowiedzialności dyscyplinarnej za naruszenie przepisów obowiązujących w uczelni oraz za czyn uchybiający godności studenta.”, oświadczam, że niniejszą pracę dyplomową wykonałem(-am) osobiście i samodzielnie i nie korzystałem(-am) ze źródeł innych niż wymienione w pracy.

Jednocześnie Uczelnia informuje, że zgodnie z art. 15a ww. ustawy o prawie autorskim i prawach pokrewnych Uczelni przysługuje pierwszeństwo w opublikowaniu pracy dyplomowej studenta. Jeżeli Uczelnia nie opublikowała pracy dyplomowej w terminie 6 miesięcy od dnia jej obrony, autor może ją opublikować, chyba że praca jest częścią utworu zbiorowego. Ponadto Uczelnia jako podmiot, o którym mowa w art. 7 ust. 1 pkt 1 ustawy z dnia 20 lipca 2018 r. — Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2018 r. poz. 1668 z późn. zm.), może korzystać bez wynagrodzenia i bez konieczności uzyskania zgody autora z utworu stworzonego przez studenta w wyniku wykonywania obowiązków związanych z odbywaniem studiów, udostepniając utwór ministrowi właściwemu do spraw szkolnictwa wyższego i nauki oraz korzystać z utworów znajdujących się w prowadzonych przez niego bazach danych, w celu sprawdzania z wykorzystaniem systemu antyplagiatowego. Minister właściwy do spraw szkolnictwa wyższego i nauki może korzystać z prac dyplomowych znajdujących się w prowadzonych przez niego bazach danych w zakresie niezbędnym do zapewnienia prawidłowego utrzymania i rozwoju tych baz oraz współpracujących z nimi systemów informatycznych.

1. Wstęp

Obecność człowieka na Ziemi dzieli się na epoki, nazwane od najważniejszych materiałów używanych do dalszego rozwoju naszego gatunku. Była epoka kamienia łupanego, epoka brązu, a z dziejów mniej odległych - epoka pary. Według tej klasyfikacji, drugą dekadę XXI wieku możemy bez wątpienia nazwać epoką **informacji**. Algorytmy do analizy danych nigdy nie były tak powszechnie używane jak obecnie. Szacuje się, że generowane codziennie ilości danych przekraczają swoją licznością wszystkie informacje zebrane przez ludzkość od początku jej istnienia do momentu powstania Internetu. Ciężko więc sobie wyobrazić, że w cieniu tego bezmiaru wciąż istnieją obszary nauki, w których rejestrowanych danych jest stanowczo za mało. Jednym z tych obszarów jest fizyka cząstek elementarnych.

Sam fakt dostępu do dużych zbiorów danych to oczywiście nie wszystko - aby wyciągnąć wnioski konieczna jest skomplikowana analiza tych informacji. Dziedziną nauki doskonale wpasowującą się w potrzeby współczesnego, skoncentrowanego na danych świata jest Uczenie Maszynowe (ang. Machine Learning). Technika ta, choć powszechnie wykorzystywana w wielu zagadnieniach z pogranicza informatyki i statystyki, w fizyce cząstek wysokich energii została zauważona dopiero kilka lat temu, a obecnie przechodzi fazę dynamicznego rozwoju. Jedną z najczęstszych aplikacji algorytmów Uczenia Maszynowego, również w fizyce, jest zagadnienie klasyfikacji, czyli określenie przynależności elementów zbioru danych do klas na podstawie zaobserwowanych wartości atrybutów klasyfikowanych obiektów.

Badania prowadzone w ramach fizyki cząstek elementarnych są niesamowicie kosztowne i czasochłonne - niezbędne jest wykorzystanie najnowszych technologii, budowa gigantycznych detektorów, oraz zaangażowanie najświatlejszych umysłów. Urządzeniem pełniącym w ostatnich latach kluczową rolę w badaniach nad rzadką i nie występującą na Ziemi materią jest niewątpliwie Wielki Zderzacz Hadronów (ang. LHC - Large Hadron Collider).

Wśród wielu nowych obszarów badań powstałych w wyniku działania urządzeń pokroju LHC wyróżnić można badania nad rzadkimi rozpadami hadronowymi. Analiza tych rozpadów jest trudna, z uwagi na niewielkie ilości rejestrowanych przypadków.

Z tego samego powodu, próby ich klasyfikacji przy użyciu technik Uczenia Maszynowego nie osiągają optymalnej, znanej z innych dziedzin nauki dokładności. Niniejsza praca skupia się na zagadnieniu poprawy jakości klasyfikacji rzadkiego rozpadu $B_s \to D_s^* K^*$ poprzez wykorzystanie odnoszącej duże sukcesy w innych dziedzinach Uczenia Maszynowego techniki powielania zbioru danych (ang. data augmentation).

Praca została podzielona na następujące części: w rozdziale drugim przedstawiona została idea technik Uczenia Maszynowego, przedstawienie języka programowania Python, opis akceleratora LHC, charakterystyka eksperymentu LHCb (ang. Large Hadron Collider beauty), oraz omówienie analizowanego kanału rozpadu. Trzeci rozdział skupia się na metodach zbierania i preselekcji danych (ang. data flow) w eksperymencie LHCb. Rozdział czwarty to omówienie wybranych metod zaawansowanej analizy danych. W rozdziale piątym przedstawiono szczegółowo tematykę pracy - algorytm, trudności, oraz potencjalne korzyści płynące z powielania danych w kontekście klasyfikacji metodą Wzmacnianych Drzew Decyzyjnych (ang. BDT - Boosted Decision Trees). Omówienie uzyskanych wyników i płynące z nich wnioski przedstawia rozdział szósty.

2. Wstęp teoretyczny

2.1. Uczenie Maszynowe

Od początków rozwoju informatyki, człowiekowi towarzyszyło marzenie o stworzeniu maszyny zdolnej do samodzielnego uczenia się i podejmowania świadomych decyzji. Zadanie okazało się być na tyle trudnym, że do dziś nie udało się stworzyć systemu spełniającego definicję "Sztucznej Inteligencji". Choć definicji tych jest kilka, ta ukuta przez Johna McCarthy'ego - autora wyrażenia "sztuczna inteligencja" - w 1956 roku daje dobry obraz oczekiwań stawianych przed inteligentną maszyną: "zdolność systemu do prawidłowego interpretowania danych pochodzących z zewnątrz, nauki na ich podstawie oraz wykorzystywania tej wiedzy, aby wykonywać określone zadania i osiągać cele poprzez elastyczne dostosowanie" [1]. Próby realizacji tego marzenia dały jednak podstawy do rozwoju szeregu algorytmów, zwanych obecnie technikami Uczenia Maszynowego, których celem jest tworzenie automatycznych systemów zdolnych do doskonalenia swojego działania przy pomocy zgromadzonego doświadczenia (danych). Z uwagi na fakt, że systemy te trenowane są na zbiorach danych, zbiory te powinny być odpowiednio duże.

Systemy Uczenia Maszynowego znajdują obecnie szerokie zastosowanie w wielu dziedzinach nauki - począwszy od analizy obrazów, poprzez algorytmy rozpoznawania mowy, automatyczne sterowanie pojazdami, na całym szeregu rozwiązań z zakresu eksploracji danych (ang. data mining) skończywszy.

Uczenie Maszynowe znalazło zastosowanie również w fizyce wysokich energii. Prócz omawianych w czwartym rozdziale niniejszej pracy technik analizy wielowymiarowej (ang. MVA - Multivariate Analysis) ze szczególnym uwzględnieniem techniki Wzmacnianych Drzew Decyzyjnych w poszukiwaniach rzadkich rozpadów mezonów pięknych i powabnych, algorytmy Uczenia Maszynowego używane są m.in w rekonstrukcji śladów cząstek w detektorach, a także w systemach wyzwalania zapisu przypadków – triggerach.

2.2. Język Python w analizie danych

Jeszcze kilka lat temu dominacja kompilowanych języków programowania w dziedzinach analizy danych była bezdyskusyjna. Języki takie jak C czy C++, a wcześniej Fortran, były wybierane przez naukowców z uwagi na ich wysoką wydajność w przetwarzaniu dużych ilości informacji. Ceną za wydajność programu była jednak stosunkowo skomplikowana składnia, oraz narzut czasowy niezbędny programiście do opanowania języka. Od kilku lat trend ten ulega zmianie. Odsetek języków kompilowanych w zagadnieniach statystyki i eksploracji danych znacznie zmalał na rzecz języków skryptowych, takich jak Python. [2]

Głównym zarzutem przez lata stawianym językom skryptowym była ich niewystarczająca wydajność w zagadnieniach wymagających dużych mocy obliczeniowych. Obecnie można z całą pewnością stwierdzić, że w zdecydowanej większości aplikacji wydajność języka Python jest odpowiednia do stawianych przed nim zadań. Fakt, że najpopularniejsza implementacja Pythona wykorzystuje język C (CPython), pozwala na połączenie szybkości działania kodu języka C z czytelnym, spójnym, i łatwym do opanowania interfejsem programisty języka Python. Dzięki takiemu połączeniu uzyskuje się wydajne narzędzie, którego opanowanie nie wymaga setek godzin nauki.

2.3. Biblioteka XGBoost w Uczeniu Maszynowym

XGBoost [3] to obecnie jeden z najpopularniejszych algorytmów uczenia maszynowego, wykorzystywany zarówno w aplikacjach regresyjnych, jak i klasyfikacyjnych. Jego główną zaletą jest szybkość działania (oryginalna implementacja algorytmu napisana została w języku C++) i możliwość zrównoleglenia wykonywanych operacji. W dzisiejszych czasach - czasach masowego wykorzystania procesorów graficznych (ang. GPU - Graphics Processing Unit) do skomplikowanych operacji, możliwość wsparcia obliczeń przez koprocesory jest nieodłączną cechą dobrego algorytmu Uczenia Maszynowego.

Z uwagi na wymienione wyżej zalety algorytmu XGBoost, oczywistym staje się fakt jego wtórnej implementacji w językach wykorzystywanych do obróbki danych, jak R czy Python. W niniejszej pracy zastosowanie znalazła biblioteka XGBoost języka Python [4] którego wykorzystanie zostało uzasadnione w poprzednim podrozdziale.

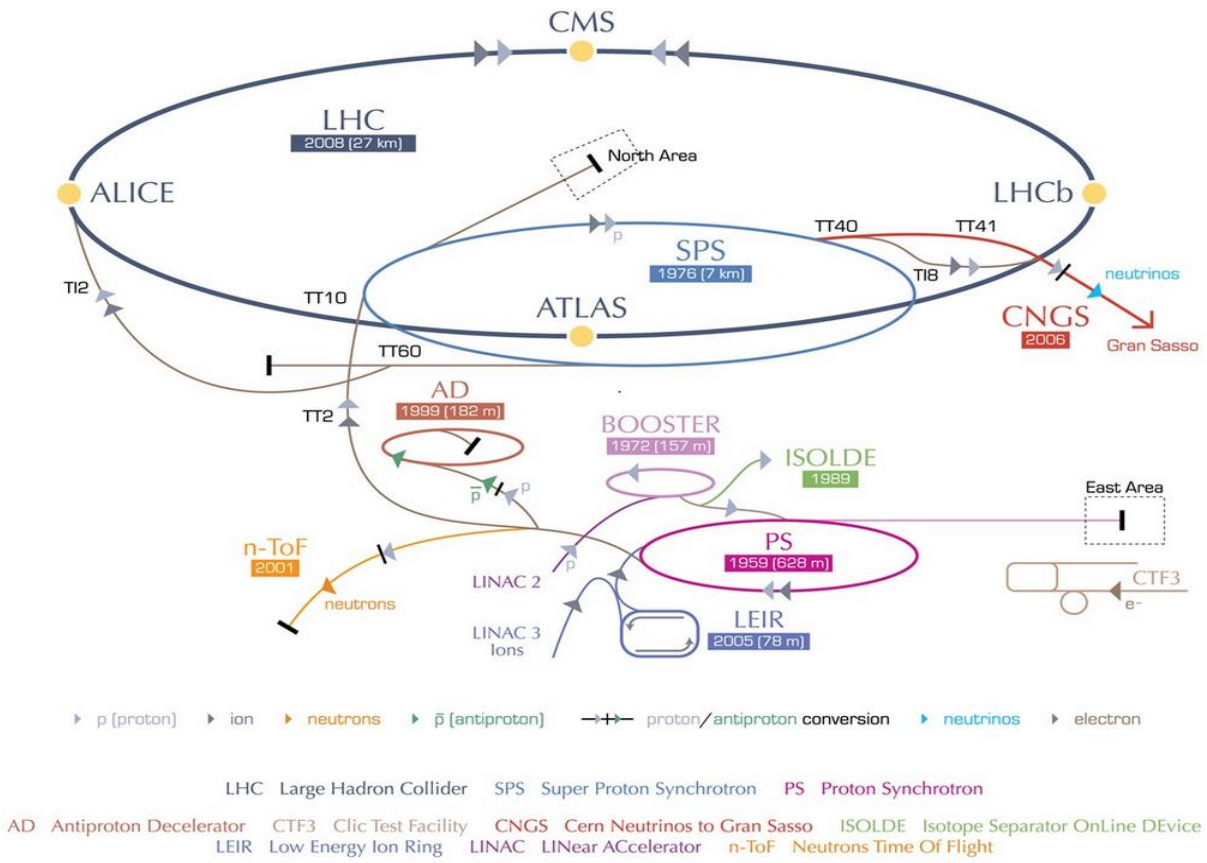
2.4. LHC

Wielki Zderzacz Hadronów to największy akcelerator cząstek na świecie [5]. Znajduje się on w Europejskim Ośrodku Badań Jądrowych (CERN) pod Genewą, i jest wynikiem finansowej i naukowej kooperacji wielu państw. Głównymi obszarami badań prowadzonych w CERN są: fizyka wysokich energii (z wykorzystaniem LHC i mniejszych akceleratorów), fizyka jądrowa, fizyka medyczna, informatyka, i inne. 27-kilometrowy torus tworzący Wielki Zderzacz leży około 100 metrów pod ziemią, na terenie Francji i Szwajcarii, a jego budowa, rozpoczęta w 2000 roku, trwała 8 lat. Pierwsze uruchomienie akceleratora zostało gwałtownie zakończone przez wybuch powstały w wyniku wycieku ciekłego helu (używanego jako chłodziwa), połączonego ze zwarciem. Naprawa skutków tego zdarzenia trwała 14 miesięcy, a do ponownego uruchomienia doszło w 2009 roku. [6]

Ogromne wymiary urządzenia zapewniają wysoką świetlność (ang. luminosity) - parametr określający liczbę zderzeń cząstek rozpędzanych w akceleratorze - oraz energię wiązki, która decyduje o możliwych stanach końcowych powstałych w wyniku kolizji hadronów. Większa energia zderzenia pozwala na powstanie (i późniejszą obserwację) masywniejszych cząstek. Liczba powstających w eksperymencie cząstek wynika bezpośrednio z liczby kolizji (a więc świetlności). Ze wzrostem tej wartości wzrasta również szansa na zaobserwowanie rozpadów rzadkich.

W celu rozpędzenia cząstek do ich docelowej energii wykorzystuje się akceleratory kołowe i liniowe znajdujące się w otoczeniu LHC. W przeciwieństwie do swojego poprzednika - LEP (zderzacza elektronów), w LHC zderzane są *hadrony*, czyli protony lub ciężkie jony. Korzyścią płynącą z tego faktu jest mniejsza utrata energii w trakcie zakrzywionego lotu cząstki wynikająca z promieniowania hamowania (niem.

Bremsstrahlung), co przekłada się na bardziej efektywną pracę akceleratora. Schemat kompleksu akceleratora LHC wraz z wchodzącymi w jego skład eksperymentami przedstawia rysunek 1.



Rysunek 1 - Schemat kompleksu akceleratorów działających w ramach LHC [7]

Wśród największych eksperymentów działających w ramach LHC wyróżnić można:

- ATLAS (ang. A Toroidal LHC Apparatus) - z założenia jest detektorem ogólnego przeznczenia, mającym za zadanie detekcję możliwie największego spektrum cząstek powstałych w wyniku zderzeń. Wykorzystywany jest również do poszukiwania i badania właściwości bozonu Higgsa oraz do sprawdzania Teorii Supersymetrii.
- ALICE (ang. A Large Ion Collider Experiment) - przeznaczony przede wszystkim do obserwacji zderzeń ciężkich jonów, w których może powstać plazma kwarkowo - gluonowa.

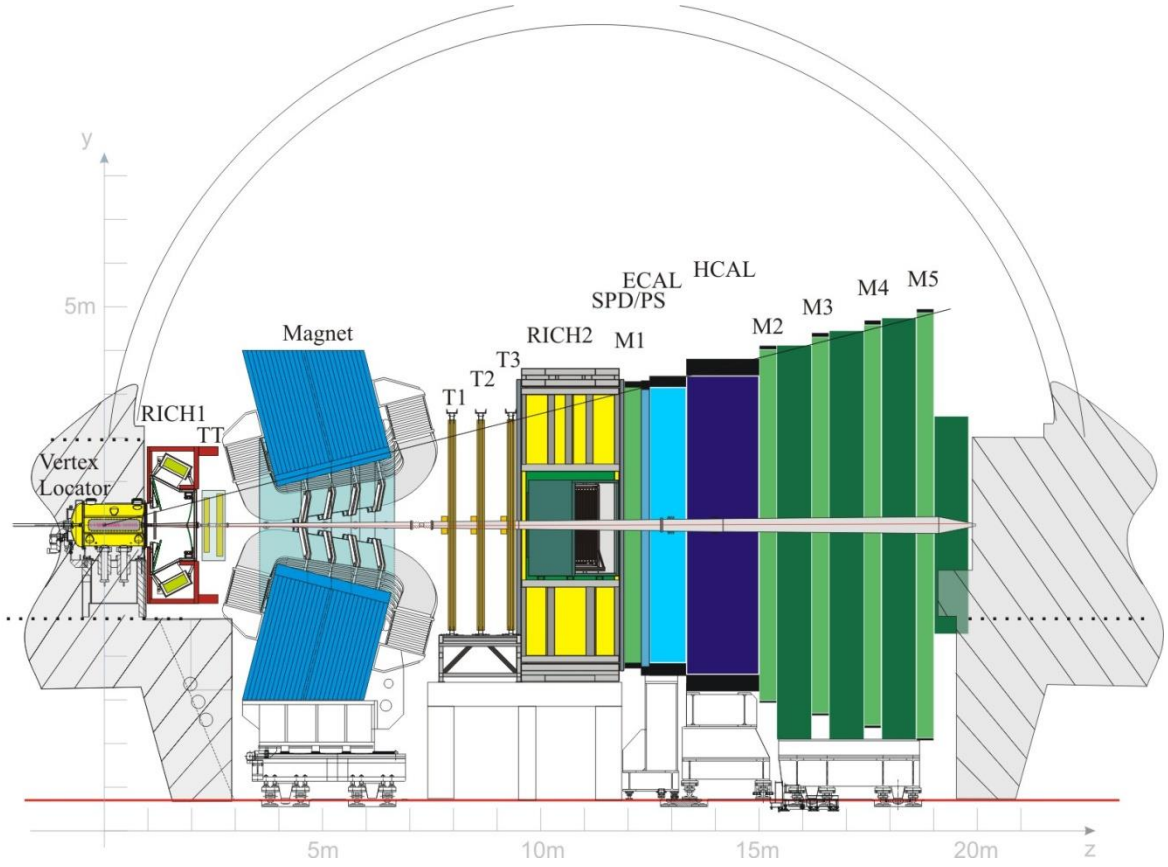
- CMS (ang. Compact Muon Solenoid) - zakres badań jest zbliżony do eksperymentu ATLAS, jednak wykorzystywany detektor cechuje się inną geometrią, co umożliwia niezależną weryfikację uzyskiwanych wyników.
- LHCb - badający fizykę ciężkich zapachów *b* i *c*.

Niniejsza praca opiera się na danych otrzymywanych ze spektrometru LHCb.

Kalendarz LHC składa się z trwających kilka lat cykli zbierania danych (ang. Run), oddzielonych od siebie również kilkuletnimi okresami wyłączenia (ang. Long Shutdown). Obecnie, od 2018 roku, akcelerator znajduje się w fazie Long Shutdown II. Czas ten poświęcany jest na prace konserwacyjne i wprowadzanie ulepszeń w całym kompleksie akceleratora. Zakończenie tego okresu planowane jest na 2022 rok. [8]

2.5. Large Hadron Collider beauty - LHCb

Spektrometr LHCb jest jednym z siedmiu detektorów zainstalowanych na Wielkim Zderzaczu Hadronów. Jego głównym przeznaczeniem jest badanie rozpadów ciężkch hadronów zawierających kwarki b i c. Z uwagi na charakterystykę tych rozpadów, rejestruje on cząstki poruszające się pod małymi kątami względem osi wiązki przyspieszanej w LHC (-300 do 300 mrad w osi pionowej, -250 do 250 mrad w osi poziomej). Detektor LHCb składa się z szeregu systemów (poddetektorów), z których każdy przeznaczony jest do pomiaru pewnej charakterystycznej właściwości cząstek produkowanych w wyniku zderzeń proton-proton (pęd, energia, identyfikacja). Rysunek 2 przedstawia schemat przekroju poprzecznego detektora.



Rysunek 2 - Schemat przekroju poprzecznego spektrometru LHCb [9]

Najważniejsze widoczne na schemacie elementy spektrometru LHCb to:

- VELO (ang. Vertex Locator) - krzemowy detektor znajdujący się najbliżej miejsca zderzeń hadronów. Jego zadaniem jest rekonstrukcja pierwotnego wierzchołka oddziaływania proton-proton oraz wierzchołków wtórnych, a także pomiar odległości między nimi. Dane zebrane przez VELO są kluczowe w badaniach ciężkich hadronów.
- Detektory Czerenkowa (ang. Ring Imaging Cherenkov) - RICH1, położony przed magnesami, przeznaczony do identyfikacji naładowanych cząstek hadronowych o niskich pędach, oraz RICH2, znajdujący się za magnesami, identyfikujący cząstki o wysokich pędach.
- Detektory śladowe: TT (ang. Track Turicensis), T_1, T_2, T_3 - oddzielone wielkim elektromagnesem, którego zadaniem jest zakrzywienie toru lotu cząstek, co pozwala na pomiar ich pędu i ładunku elektrycznego. Detektory znajdujące się

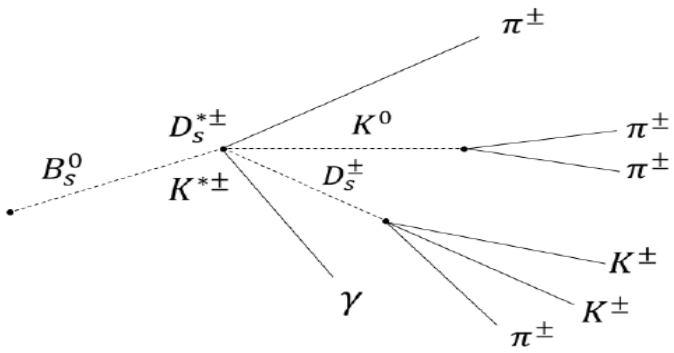
za magnesem - T_1, T_2, T_3 składają się z części krzemowej (blisko osi głównej akceleratora), oraz gazowej (zewnętrznej).

- Kalorymetry ECAL i HCAL (ang. Electromagnetic Calorimeter, Hadronic Calorimeter) - absorbują cząstki, mierząc w efekcie ich energię
- Detektory mionów - znajdują się za osłonami pochłaniającymi. Miony, z uwagi na ich bardzo słabe oddziaływanie z materią, są jedynymi cząstkami rejestrowanymi za osłonami.

Taka mnogość systemów składowych skutkuje produkcją ogromnych ilości informacji. Z tego powodu, rozdział trzeci został w całości poświęcony charakterystyce przepływu danych w eksperymencie LHCb.

2.6. Charakterystyka rozpadu $B_S^0 \to K^{*\pm} D_s^{*\pm}$

Rozpad $B_S^0 \to K^{*\pm} D_s^{*\pm}$ jest rzadkim i dotąd niezaobserwowanym rozpadem hadronowym mezonu B. W jego przebiegu można wyróżnić cztery stany pośrednie: $D_S^{*\pm}$; $K_S^{*\pm}$; $D_S^{\pm}$; K_S^0 , oraz siedem cząstek w stanie końcowym: foton, dwa kaony, oraz cztery piony. Stany pośrednie $D_s^{*\pm}(2112)$ i $K^{*\pm}(892)$ są tak zwanymi cząstkami rezonansowymi, charakteryzującymi się bardzo krótkim czasem życia. W praktyce oznacza to przyjęcie założenia, że rozpadają się one w tym samym punkcie, w którym powstały. Wydajność rekonstrukcji cząstek neutralnych, jak mezon K_S^0 jest niska, co skutkuje ogromną liczbą przypadków tła. Przypadki te dzielą się na tło *fizyczne* - powstające w wyniku rozpadów podobnych do poszukiwanego, które mogą przejśc przez kryteria selekcji i kontrybuować w otrzymanych rezultatach, oraz *kombinatoryczne* - wynikające z błędnej rekonstrukcji rozpadu. Liczność tła sprawia, że wydobycie zeń sygnału jest zadaniem skomplikowanym. Schemat omawianego rozpadu został przedstawiony na rysunku 3.



Rysunek 3 - Diagram rozpadu $B_S^0 \rightarrow K^{*\pm} D_s^{*\pm}$

3. Eksperyment LHCb - przepływ danych (ang. data flow)

W ciągu sekundy, w akcelatorze LHC dochodzi do 40 000 000 zderzeń wiązek proton-proton. W efekcie, ilość danych zbieranych w systemach działających w ramach LHCb sięga 1 TB/s. Takie ilości informacji nie mogą być w czasie rzeczywistym zapisane do pamięci masowej. Do ograniczenia rozmiaru zapisywanych danych wykorzystywany jest system wyzwalania przypadku - tryger (ang. trigger). Zadaniem trygera jest analiza rozpadów zbieranych przez poddetektory, ocena czy dany rekord ukazuje warty uwagi proces fizyczny, i czy w konsekwencji powinien zostać zrekonstruowany i zapisany, czy też pominięty.

W trakcie okresów zbierania danych Run I oraz Run II tryger składał się z dwóch części - programowej: HLT1, HLT2 (ang. High Level Trigger), oraz sprzętowej: L0. Komponent HLT zaimplementowany jest na farmie obliczeniowej. Skutkiem działania trygera jest zmniejszenie strumienia danych z ~40 MHz do ~12.5 kHz.

Kolejnym etapem procesu zbierania danych jest stripping, czyli wstępne selekcjonowanie przypadków i ich podział na kategorie, dokonywany na podstawie zestawu kryteriów tworzonego osobno dla każdego typu rozpadu.

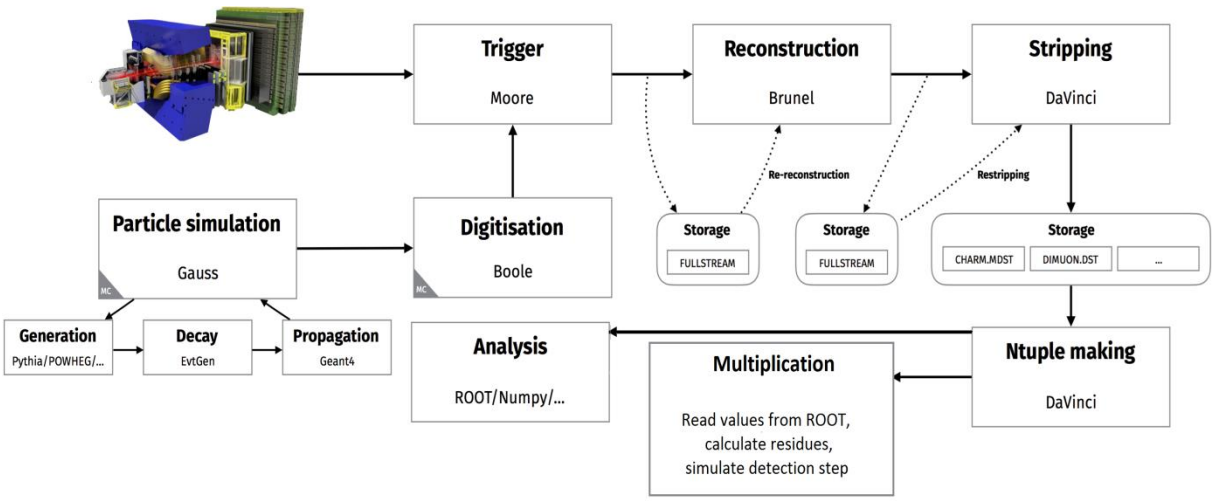
Ostatecznie, wyselekcjonowane rekordy zapisywane są w formie n-krotki (ang. n-tuple) zwanej *drzewem* (ang. tree). Ta przypominająca bazę danych struktura jest zoptymalizowana pod kątem operacji takich jak iteracja, przeszukiwanie, łączenie, czy modyfikacja jej elementów. W efekcie, czas wykonania w.w. operacji jest znacznie skrócony względem plików w innych formatach danych. Gałęziami tworzącymi drzewo są poszczególne zarejestrowane wielkości fizyczne cząstek występujących w rozpadzie.

Szybkość przeprowadzania operacji na strukturze danych to jednak tylko jeden z aspektów analizy i eksploracji danych. Drugim, równie ważnym, jest czas potrzebny człowiekowi do napisania kodu programu wykonującego tę analizę. Bez dobrego, intuicyjnego narzędzia wyspecjalizowanego w obsłudze n-krotek, zbiory te byłyby dostępne jedynie dla wąskiego grona największych specjalistów. W celu ułatwienia pracy z drzewami, w CERN stworzona została oparta na języku C++ platforma ROOT. Umożliwia ona sprawną obróbkę danych zawartych w krotkach na wszystkich

etapach analizy. Narzędzie daje możliwość odczytu, modyfikacji, wizualizacji w postaci wykresu, histogramu, mapy cieplnej i innych , a następnie zapisu rekordów w nowej krotce.

W ramach eksperymentu LHCb działają również generatory danych symulowanych, których zadaniem jest produkcja przypadków żądanego rozpadu jak najwierniej odwzorowujących strukturę danych rzeczywistych. Przypadki te poddawane są procesom ściśle naśladującym realne warunki działania systemów detektora LHCb. Niestety, w przypadku rzadkich rozpadów liczba wygenerowanych rekordów jest z reguły bardzo mała, a sama generacja wymaga ogromnej mocy obliczeniowej oraz dużych nakładów czasowych.

Niniejsza praca wprowadza pewną modyfikację do przedstawionego powyżej algorytmu przepływu danych - przed analizą uzyskanych n-krotek, zawarte w nich rekordy są powielane. Autorski algorytm multiplikacyjny został szczegółowo omówiony w 5. rozdziale tej pracy. Rysunek 4 przedstawia schemat przepływu danych w eksperymencie LHCb, uzupełniony o w.w. krok multiplikacji.



Rysunek 4 - Schemat przepływu danych w eksperymencie LHCb [10]. Na schemacie widoczny autorski krok multiplikacji danych.

4. Metody zaawansowanej analizy danych

4.1. Analiza wielu zmiennych - MVA

Metody analizy wielu zmiennych (ang. MVA - MultiVariate analysis) znajdują zastosowanie w przypadku obróbki wielowymiarowych zbiorów danych. W standardowym, jednowymiarowym podejściu do analizy, poszczególne cechy elementów takiego zbioru analizowane są osobno, w oderwaniu od kontekstu wynikającego ze zmienności innych cech i korelacji między nimi. Dzięki jednoczesnej interpretacji wielu atrybutów obiektu, MVA nierzadko dostarcza zupełnie nowych obserwacji, których nie dałoby się uzyskać rozpatrując tylko jedną cechę naraz. Jedną z najczęstszych aplikacji technik analizy wielowymiarowej jest klasyfikacja danych, czyli podział obiektów na klasy wynikające z wartości posiadanych przez nie atrybutów. Programy klasyfikujące (klasyfikatory) częstokroć tworzone są w oparciu o algorytmy Uczenia Maszynowego, i jako takie, mogą być podzielone na dwie grupy: oparte na uczeniu nadzorowanym (ang. supervised learning), oraz oparte na uczeniu nienadzorowanym (ang. unsupervised learning).

Uczenie nadzorowane jest to zbiór technik zakładających ludzki nadzór nad procesem odwzorowania wejścia programu na jego wyjście. Nadzór ten najczęściej objawia się poprzez dostarczenie programowi zbioru danych uczących rozszerzonego o informację o prawidłowej klasie przynależnej danemu rekordowi. Istotą uczenia nienadzorowanego jest z kolei, aby program sam, bez dodatkowych informacji z zewnątrz, odkrywał istniejące w zbiorze danych wzorce. W technikach tych program nie posiada informacji o oczekiwanym (prawidłowym) wyniku operacji na danym rekordzie.

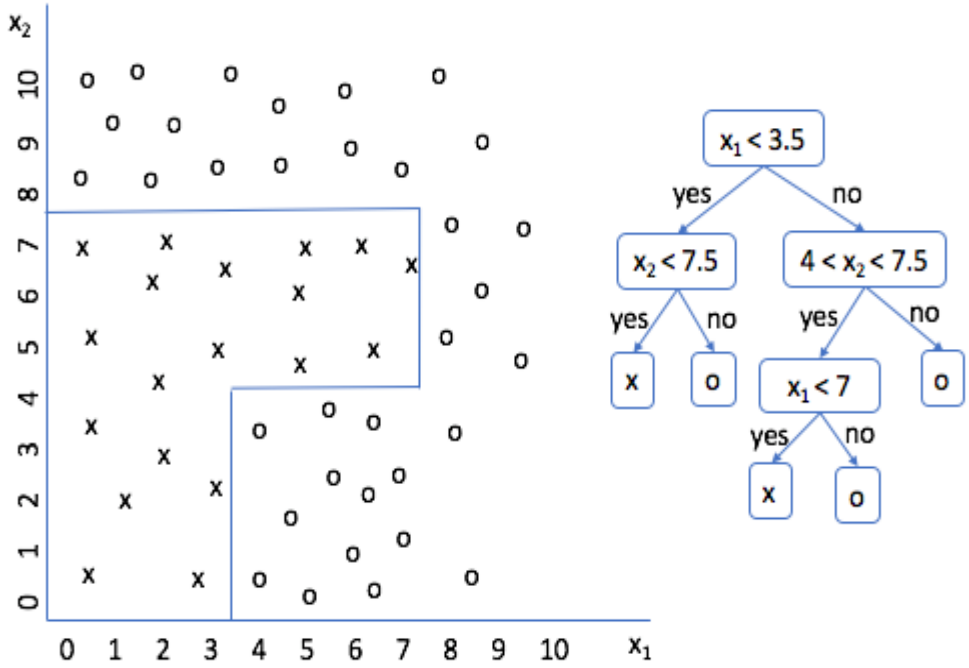
Algorytm analizy wielowymiarowej wykorzystany w niniejszej pracy do stworzenia klasyfikatora oparty jest o uczenie nadzorowane. Klasyfikator tego typu wymaga *wytrenowania*, to znaczy stworzenia uogólnionego modelu predykcyjnego, będącego w stanie sklasyfikować nowy, wcześniej nieznany obiekt na podstawie uprzedniej analizy podobnych rekordów. Celem zapewnienia wysokiej dokładności programu, zbiór przeznaczony do *trenowania* klasyfikatora dzielony jest na dwie części - uczącą oraz testową. Funkcja klasyfikująca wykorzystuje zbiór uczący do

wyszukania wzorców odpowiedzialnych za przynależność obiektu do danej klasy i stworzenia modelu predykcyjnego, zaś do sprawdzenia jego dokładności wykorzystywany jest zbiór testowy. Zastosowanie zbioru testowego pozwala stwierdzić m.in. przetrenowanie, czyli sytuację, w której model świetnie radzi sobie z klasyfikacją zbioru uczącego, ale nie jest zdolny do uogólnienia wykorzystywanych zależności w celu poprawnej klasyfikacji nowych, wcześniej nie widzianych przypadków.

Analiza wielu zmiennych dobrze wpasowuje się w potrzeby fizyki wysokich energii - cząstki powstające w wyniku zderzeń w akceleratorach opisywane są wieloma parametrami, których wzajemne zależności nie mogą być pominięte. Dzięki zastosowaniu metod MVA możliwe jest zautomatyzowanie procesu wykrywania przypadków sygnałowych, oraz odrzucenie przypadków stanowiących tło.

4.2. Drzewa decyzyjne

Drzewo decyzyjne (ang. Decision Tree) to algorytm klasyfikujący oparty o nadzorowane uczenie maszynowe, polegający na podziale przestrzeni zmiennych na obszary, których krawędzie są równoległe do osi współrzędnych. Każdemu obszarowi odpowiada pewna klasa, a klasyfikacja sprowadza się do przyporządkowania obiektu do tej klasy, w granicach której plasują go wartości jego cech (współrzędnych). Czytelniejszym sposobem przedstawienia tej techniki jest wykorzystanie struktury drzewa, w którym węzły reprezentują decyzje (test na zmiennych danego wektora), zaś gałęzie są możliwymi wynikami testu. Końcowe gałęzie drzewa prowadzą do liści, będących klasami do których może przynależeć obiekt. Ta forma przedstawienia drzewa decyzyjnego znacznie ułatwia interpretację - poszczególne testy i ich konsekwencje są wyraźnie widoczne, nawet dla średnio wprawnego obserwatora. Możliwa jest również łatwa analiza jak należałoby zmienić wartości cech obiektu, aby został on zakwalifikowany inaczej. Schemat prostego klasyfikatora opartego na drzewie decyzyjnym przedstawia rysunek 5.



Rysunek 5 - Schemat klasyfikacji metodą drzewa decyzyjnego [11]

Tworząc drzewo decyzyjne należy znaleźć kompromis między efektywnością klasyfikacji a prostotą struktury. Jedną z metryk określającą jakość klasyfikacji (a co za tym idzie - jakość dopasowania parametrów modelu) jest indeks Gini, określony wzorem:

$$Q_T(T) = \sum_{k=1}^{K} p_{rk} \cdot (1 - p_{rk})$$

gdzie p_{rk} określa jaka część zbioru danych została przypisana do klasy k. Kryterium oceny stopnia rozbudowy drzewa jest tutaj liczba przypadków przyporządkowanch do jednego liścia (po przekroczeniu progu nowe gałęzie nie są dodawane), lub funkcja straty.

Inną metodą optymalizacji jest przycinanie drzewa. Celem stosowania tej techniki jest uproszczenie drzewa, przy jednoczesnym zachowaniu wysokiej jakości odwzorowania. Korzyścią płynącą z przycięcia drzewa jest zazwyczaj oszczędność pamięci obliczeniowej, oraz skrócenie czasu klasyfikacji. Wśród algorytmów stosowanych w tej technice optymalizacji można wyróżnić:

- Expected error pruning - polegający na rekursywnym usunięciu gałęzi, dla których funkcja błędu poprzedniego rozgałęzienia jest mniejsza od sumy błędów gałęzi poniżej
- Cost complexity pruning - polegający na porównaniu poprawy klasyfikacji całego rozgałęzienia względem efektywności klasyfikacji pojedynczego rozgałęzienia w tym samym miejscu

Metodą uczenia drzew decyzyjnych wykorzystywaną w niniejszej pracy jest boosting (ang. Boosted Decision Trees - wzmacnianie drzewa decyzyjne). W technice tej wiele słabych klasyfikatorów jest łączone w jeden mocny.

5. Powielanie danych

5.1. Motywacja

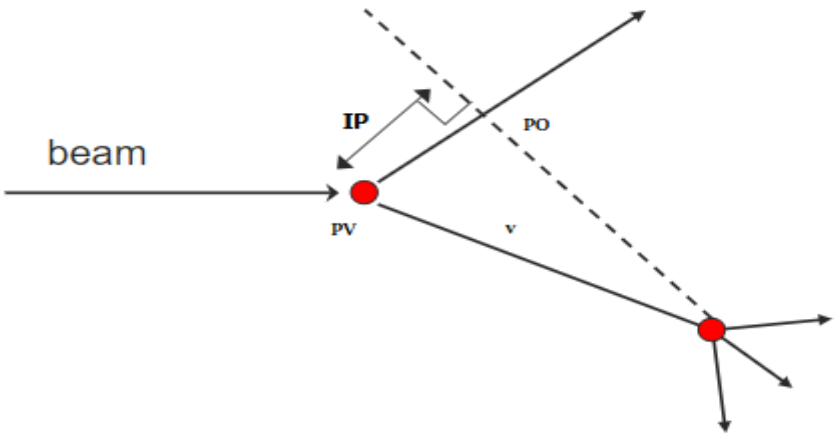
Porównując prawdopodobieństwa rozpadów podobnych kanałów rozpadu mezonu B [12], spodziewana statystyka liczby przypadków jest bardzo mała - mówimy tu o kilku tysiącach przypadków w całym okresie zbierania danych w eksperymencie LHCb. Z drugiej strony, chcąc wykorzystać klasyfikatory oparte o metody analizy wielu zmiennych, potrzeba nam zbiorów na tyle dużych, aby móc podzielić je na dane treningowe i testowe, oraz na tyle uogólnionych, aby algorytm klasyfikujący był w stanie poprawnie przenieść wyuczone reguły klasyfikacji na nowe, wcześniej nie rejestrowane przypadki. Z uwagi na powyższe, niezbędne staje się wykorzystanie danych symulowanych, używanych m.in. w trenowaniu modeli uczenia maszynowego, oraz do wstępnych badań poszukiwanych rozpadów. Niestety, z uwagi na konieczność dokładnego odwzorowania procesu rozpadu i detekcji (tak, aby wszystkie cząstki zostały poprawnie zidentyfikowane i zrekonstruowane), finalna liczba wygenerowanych przypadków Monte Carlo również jest niewystarczająca.

W ostatnich latach - latach bardzo szybkiego rozwoju technik uczenia maszynowego, podobne problemy dotknęły również inne dziedziny informatyki. Narodził się pomysł, aby przejrzeć rozwiązania wypracowane przez naukowców z innych obszarów nauki o danych, i spróbować zaadaptować je do potrzeb fizyki wysokich energii. Obiecującą techniką wydało się rozwiązanie wprowadzone w dziedzinie komputerowej analizy obrazów. W opisywanej technice, niedobory ilościowe danych są niwelowane poprzez sztuczne powielenie przypadków. W praktyce sprowadza się to do wykorzystania wycinków uprzednio wykorzystanych obrazów, lub ich nieznacznie zmodyfikowanych kopii. Metoda ta dała nadspodziewanie dobre wyniki, stąd pomysł, aby spróbować zaadaptować ją do potrzeb badania rzadkich rozpadów. Konieczne było jednak opracowanie autorskiego algorytmu multiplikującego dane, który działałby zgodnie ze znanymi prawidłami fizycznymi.

5.2. Algorytm multiplikujący

Autorski algorytm multiplikujący wykorzystany niniejszej pracy jest pierwszym lub jednym z pierwszych zastosowań techniki sztucznego powielania danych w fizyce wysokich energii. Z tego powodu szczególną uwagę poświęcono zapewnieniu zgodności nowo powstałych przypadków z rzeczywistymi wielkościami fizycznymi. Pierwszym krokiem była analiza rozpadu $B_S^0 \rightarrow K^{*\pm} D_s^{*\pm}$ i wybór zbioru parametrów, których multiplikacja da największe szanse na poprawę jakości klasyfikacji. W zbiorze tym znalazły się:

- pędy stanów końcowych, których odpowiednia modyfikacja, a następnie rekonstrukcja drzewa rozpadu pozwalały stworzyć częściowo nowe przypadki
- pędy poprzeczne - wypadkowa składowych pędu cząstki wzdłuż osi X i Y. Z uwagi na fakt, że pęd wzdłuż osi cząstki (osi Z) jest znacznie większy od pędu wzdłuż pozostałych osi, pęd poprzeczny stanowi interesującą informację
- wierzchołki powstania i rozpadu cząstek
- geometryczny parametr zderzenia (ang. Impact Parameter) - prostopadła odległość między torem lotu cząstki, a środkiem pola potencjału obiektu, do którego cząstka się zbliża (rysunek 6)



Rysunek 6 - Wyjaśnienie definicji geometrycznego parametru zderzenia

W opracowanej procedurze, produkcja nowych przypadków bazuje na modulacji efektów aparaturowych. Stworzony w toku pracy algorytm oblicza dla wybranych parametrów residuum, czyli różnicę między ich wartościami wygenerowanymi w symulowanych zderzeniach, a wartościami po etapie symulowanej odpowiedzi detektora. W założeniu, rozkład wartości residuum dla każdej cechy powinien oscylować wokół zera, z odchyleniem standardowym wielokrotnie mniejszym niż rozkład oryginalnych wartości danej cechy. Założenie to wynika z faktu, że odpowiedź detektora nie może mieć zbyt dużego wpływu na rzeczywiste wartości - wtedy pomiary byłyby obarczone dużym błędem, i nie miałyby większego sensu. Otrzymawszy rozkład wartości residuum z wielu elementów zbioru danych, wyznaczono funkcję opisującą jego parametry statystyczne, takie jak wartość średnia i odchylenie standardowe. W kolejnym kroku funkcja ta została zaimplementowana w generatorze pseudolosowym. Uzyskano w ten sposób generator odpowiedzi detektora, w założeniu reprezentujący wszystkie procesy związane z elektroniką odczytu i procesem detekcji. Podejście to pozwala na wielokrotne wykorzystanie danych, dla których znane są wielkości atrybutów przed etapem symulowanej detekcji - do wielkości tych dodajemy każdorazowo kolejną wartość z generatora odpowiedzi detektora, otrzymując nowy, wpasowujący się w ramy fizyczne przypadek.

Dalej działanie algorytmu różni się w zależności od rodzaju powielanego parametru. W przypadku pędów stanów końcowych, po etapie symulacji odpowiedzi detektora następuje pełna rekonstrukcja drzewa rozpadu bazująca na nowopowstałych stanach końcowych. Celem rekonstrukcji jest odtworzenie parametrów wszystkich cząstek - matek biorących udział w rozpadzie tak, aby parametry te uwzględniały różnicę wynikającą z modyfikacji cech stanów końcowych. Szczegółowe omówienie procesu rekonstrukcji wraz z przykładowymi wynikami przedstawiono w rozdziale 5.3. W przypadku pędów poprzecznych, nowo uzyskane wielkości są wstawiane w odpowiednie pola n-krotki i zapisywane.

Z kolei powielając wierzchołki powstania i rozpadu cząstek należy pamiętać, o kilku ważnych ograniczeniach wynikających z natury rozpadu i detektora:

- Miejsce rozpadu cząstki jest miejscem powstania cząstek - córek, w związku z czym wierzchołek rozpadu jednej cząstki jest jednocześnie wierzchołkiem powstania innej, i wstawiając jeden z nich do krotki, należy tą samą wartością wypełnić odpowiednie pole cząstki pochodnej.
- Wyjątkiem od powyższej reguły jest mezon $\boldsymbol{B_S}$, który powstaje w miejscu zderzeń proton - proton, a miejsce to nazywane jest *wierzchołkiem pierwotnym* - PV.
- W przypadku cząstek D_S^* i K^* konieczna jest inna obsługa algorytmu - są one stanami rezonansowymi, charakteryzującymi się bardzo krótkim czasem życia. Z uwagi na ograniczenia detektora w próbkowaniu położenia cząstki, przyjmuje się, że stany te rozpadają się w miejscu powstania.
- Wierzchołki rozpadu stanów końcowych nie są określone, ponieważ cząstki te nie rozpadają się w detektorze.

Jednym z bardziej skomplikowanych parametrów wykorzystywanych w analizie jest geometryczny parametr zderzenia. Do wyliczenia jego nowych wartości wykorzystywane są obliczone wcześniej rozmyte położenia wierzchołków. Dla każdej cząstki, algorytm wykorzystuje jej współrzędne wierzchołków powstania i rozpadu, oraz współrzędne wierzchołka powstania głównej cząstki (B_S^0). Z tego powodu, rozmycie *parametru zderzenia* dla niektórych cząstek nie jest możliwe - jako przykład można przedstawić stany końcowe, dla których nie posiadamy informacji o wierzchołkach rozpadu.

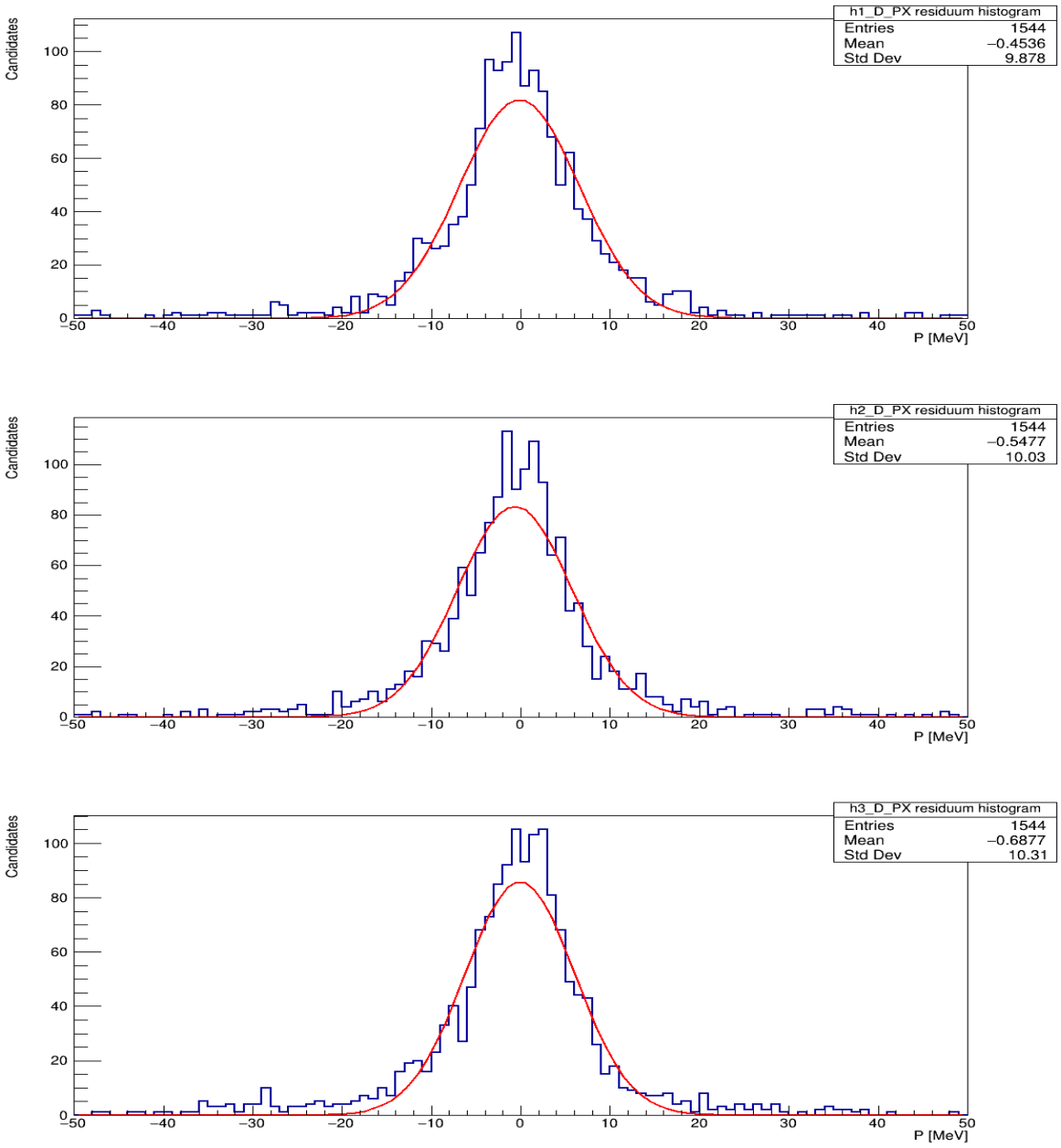
Potwierdzenie fizycznej prawidłowości powielonych przypadków odbywa się poprzez porównanie histogramów rozkładu charakterystycznych atrybutów nowych rekordów z rozkładami wynikającymi z próbki pierwotnej

5.3. Implementacja algorytmu

Jedną z pierwszych koniecznych do podjęcia decyzji w początkowych stadiach powstawania programu multiplikującego opartego o powyższy algorytm był wybór języka programowania. Z uwagi na istnienie platformy ROOT, jedynym rozsądnym

wyborem wydawał się być C++. Po dokładniejszej analizie okazało się jednak, że z uwagi na rosnącą popularność języka Python w dziedzinie obróbki danych, powstała wersja pakietu ROOT napisana z myślą o tym właśnie języku - pyROOT. Wersja ta udostępnia wszystkie funkcjonalności oryginału, oferując jednocześnie wszystkie zalety języka Python. Rozważywszy wady i zalety obydwu podejść, zdecydowano się wykorzystać implementację pyROOT.

Głównym wyzwaniem w implementacji programu multiplikującego była poprawna obsługa pochodzących z eksperymentu LHCb krotek - baz danych stosowanych w eksperymentach z zakresu fizyki wysokich energii, m.in. LHCb. Po otwarciu zawartości pliku, konieczne było powiązanie adresów zmiennych z programu z adresami odpowiadających im gałęzi drzewa zawartego w krotce. W następnym kroku program iterował po zawartości drzewa, każdorazowo przypisując zmiennym wartości z danego rekordu. Iteracja ta miała na celu obliczenie wartości residuów modyfikowanych cech dla poszczególnych przypadków. Po przeprocesowaniu wszystkich elementów drzewa, program wyliczał funkcje statystyczne residuów, a później tworzył generator symulowanej odpowiedzi detektora. Przykładowe rozkłady residuów (wykreślone dla składowej X pędu stanów końcowych h1, h2, h3) z naniesionymi krzywymi opisującymi funkcje rozkładu statystycznego przedstawia rysunek 7.

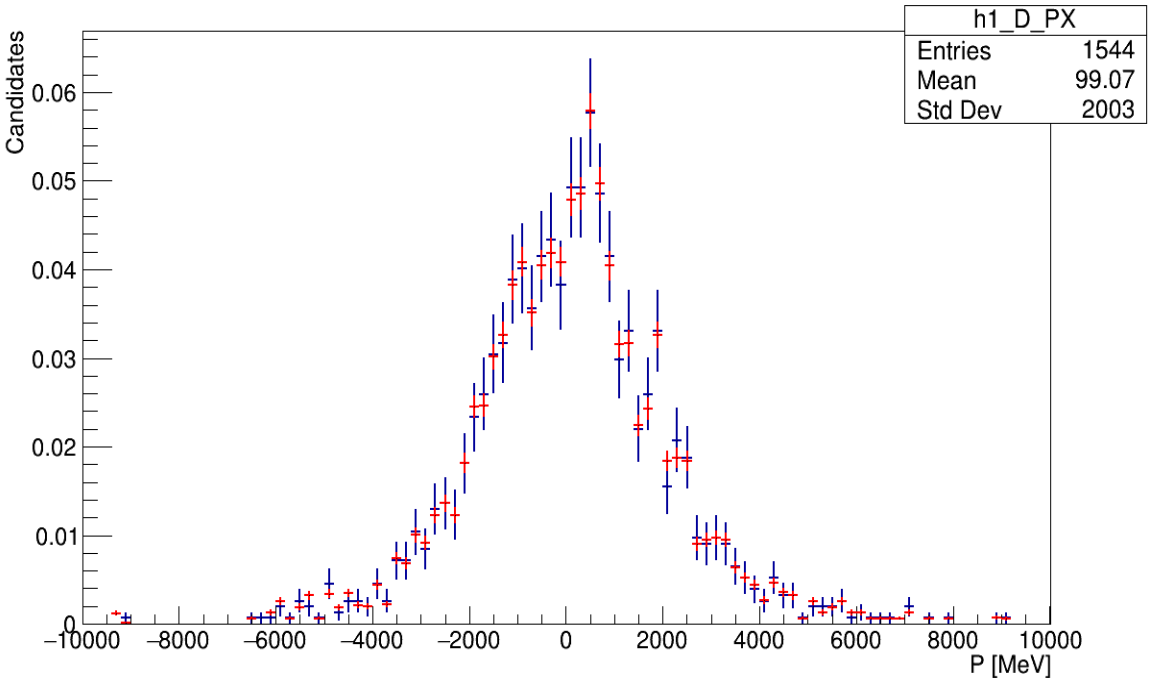


Rysunek 7 - Rozkłady przedstawiające residua obliczone dla zmienianych wartości stanów końcowych rozpadu mezonu Ds wraz z dopasowanym modelem normalnym

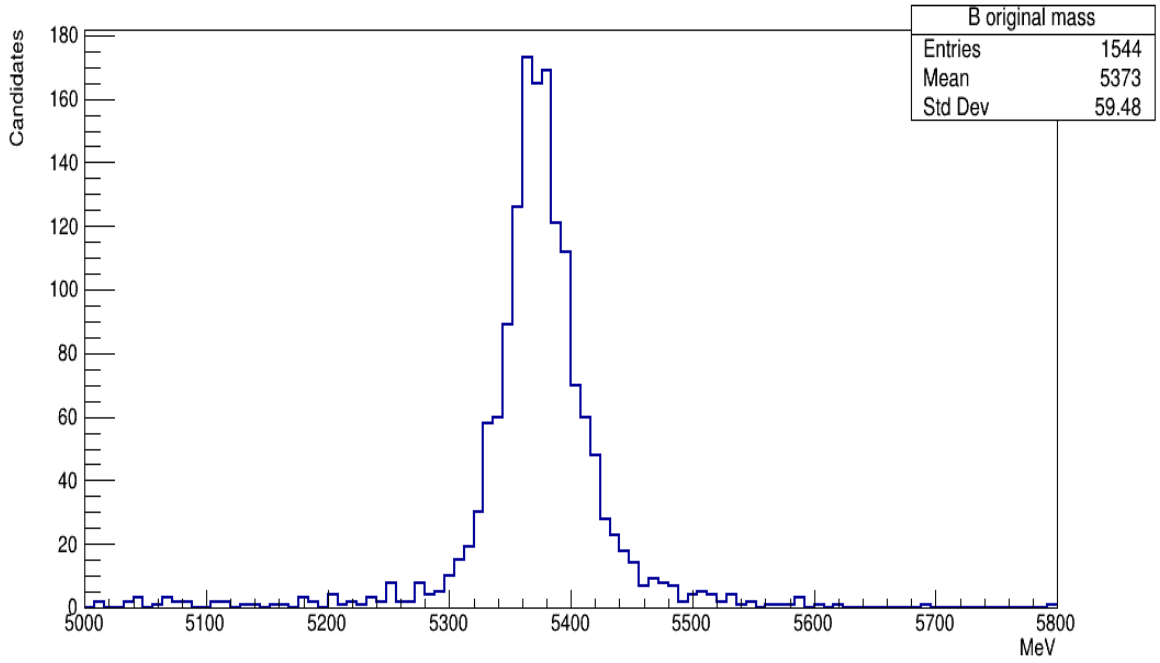
Dla każdego procesowanego przypadku rozpadu, obliczenia wykonywane były przy użyciu standardowych bibliotek i narzędzi języka Python, przy wsparciu specjalistycznych klas i metod pakietu pyROOT. Wyniki powyższych obliczeń przechowywane były w pamięci programu, bez konieczności modyfikacji n-krotki.

W kolejnym kroku rozpoczynała się nowa pętla, której zadaniem było wygenerowanie żądanej przez użytkownika liczby nowych przypadków, zgodnie z

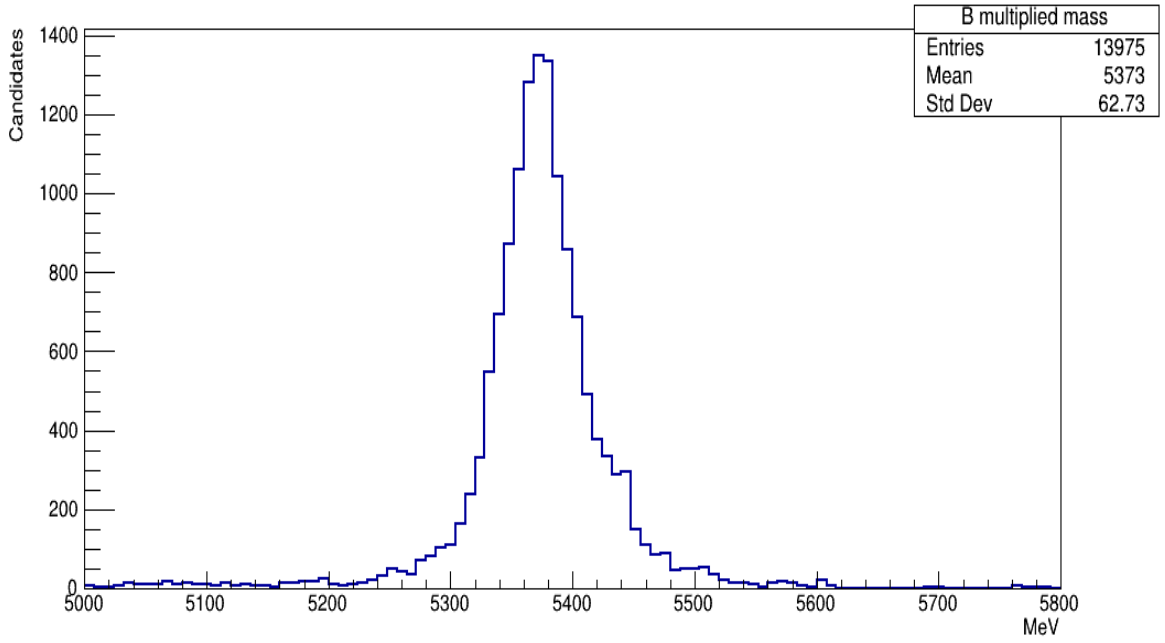
alogrytmem opisanym w punkcie 5.2. Każdy wygenerowany przypadek musiał zostać zapisany jako nowy rekord n-krotki, a wartości zmodyfikowanych atrybutów umieszczone w histogramach poszczególnych wartości charakterystycznych, w celu późniejszego sprawdzenia prawidłowości rozkładu powielonych przypadków. Sprawdzenia poprawności dokonano m.in. poprzez porównanie znormalizowanych rozkładów multiplikowanych parametrów z ich rozkładami z próbki pierwotnej (rysunek 8), oraz analizując rozkłady cech cząstek powstałych w wyniku pełnej rekonstrukcji drzewa rozpadu (rysunki 9, 10).



Rysunek 8 - Porównanie znormalizowanych rozkładów pędów wzdłuż osi X mezonów ($K^{\pm}$ / $\pi^{\pm}$) pochodzących z rozpadu mezonu Ds dla pierwszego stanu końcowego - punkty niebieskie: próbka pierwotna, punkty czerwone: próbka powielona



Rysunek 9 - Rozkład masy mezonu Bs przed multiplikacją danych



Rysunek 10 - Rozkład masy mezonu Bs po multiplikacji i rekonstrukcji

5.4. Testowanie powielonych przypadków w klasyfikatorze BDT

Dysponując powielonym zbiorem danych możliwe stało się sprawdzenie poprawy jakości klasyfikatora działającego w oparciu o technikę wzmacnianych drzew decyzyjnych. Oczekiwanym efektem jego działania było rozdzielenie dostarczonego zbioru przypadków rozpadu mezonu $\boldsymbol{B_S}$ na przypadki sygnałowe oraz tło kombinatoryczne.

W celu przetestowania działania metody powielania danych, posłużono się zbiorem rekordów zebranym w eksperymencie LHCb w 2015 i 2016 roku, o scałkowanej świetlności $\int L\,dt = 1{,}99\,fb^{-1}$. Pierwszy krok oceny poprawy jakości klasyfikacji polegał na nauczeniu dwóch osobnych klasyfikatorów - pierwszego, stworzonego na podstawie pierwotnego, nie powielonego zbioru danych, oraz drugiego, opartego o zbiór zmultiplikowany. Obydwa programy wykorzystywały ten sam zbiór zmiennych klasyfikujących, składający się z atrybutów cząstek, które udało się powielić dzięku programowi multiplikującemu, oraz innych standardowych cech używanych w analize eksperymentu LHCb:

- Pędów poprzecznych (PT) cząstek $\boldsymbol{K_S^0}$; $\boldsymbol{\gamma}$; $\boldsymbol{K^{\pm}}$ / $\boldsymbol{\pi^{\pm}}$
- Parametrów zderzenia (IP) cząstek $\boldsymbol{D}$; $\boldsymbol{K^{\pm}}$ / $\boldsymbol{\pi^{\pm}}$
- Wierzchołków powstania lub rozpadu cząstek $\boldsymbol{B}$; $\boldsymbol{D}$; $\boldsymbol{\pi^{\pm}}$

Szczegółowa lista zmiennych, wraz z ich współczynnikami *feature importance* (wielkością wpływu każdej zmiennej na efekt klasyfikacji) została omówiona w rozdziale 6.5 na rysunku 50).

Po wytrenowaniu obu programów, kolejny etap polegał na ocenie i porównaniu wyników klasyfikacji testowego zbioru rekordów. Mając do czynienia z uczeniem nadzorowanym, każdy z przypadków zaklasyfikowano do jednej z czterech grup:

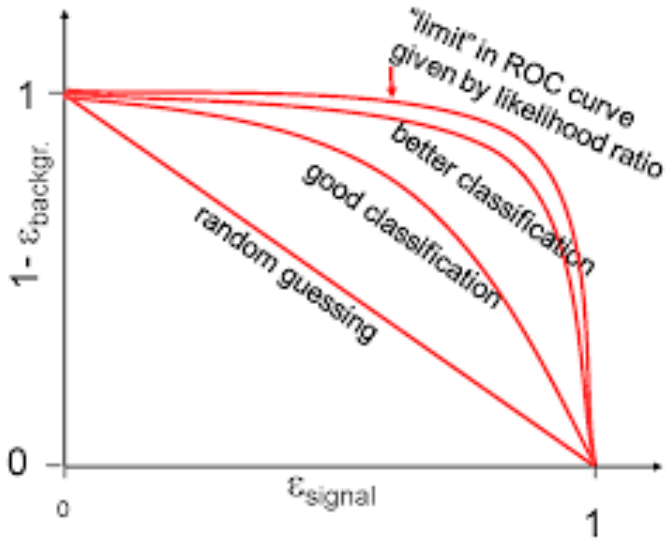
- True Positives (TP) - przypadki, w których rekord został przypisany do klasy sygnałowej, i w rzeczywistości faktycznie był to przypadek sygnałowy
- True Negatives (TN) - przypadki, w których rekord został przypisany do klasy tła, i w rzeczywistości faktycznie przypadek ten nie był sygnałem

- False Positives (FP) - przypadki, w których rekord został przypisany do klasy sygnałowej, jednak w rzeczywistości był to przypadek tła.
- False Negatives (FN) - przypadki, w których rekord został przypisany do klasy tła, jednak w rzeczywistości był to przypadek sygnałowy.

Na podstawie przedstawionych powyżej grup poddano analizie następujące wielkości charakterystyczne:

- Krzywa ROC (ang. Receiver Operating Characteristic) - wykres liczby przypadków True Positive do False Positive. Im bardziej wartość ta zbliżona jest do 1, tym lepszy klasyfikator. Przykładowe przebiegi krzywej ROC przedstawiono na rysunku 11.
- Figure of Merit (FoM) - parametr wykorzystany do oceny poprawy jakości klasyfikacji dokonanej na danych rzeczywistych. $FoM = \frac{S}{\sqrt{S+B}}$ gdzie S - liczba przypadków sygnałowych, B - liczba przypadków tła.

Liczba przypadków sygnałowych i tła została oszacowana na podstawie rozkładu teoretycznego dopasowanego do rozkładu masy $\boldsymbol{K^*(892)}$. Szczegółowy opis zastosowanej metody wraz z uzyskanymi wynikami został przedstawiony w rozdziale 6.5.



Rysunek 11 - Przebiegi krzywej ROC dla różnych jakości klasyfikacji

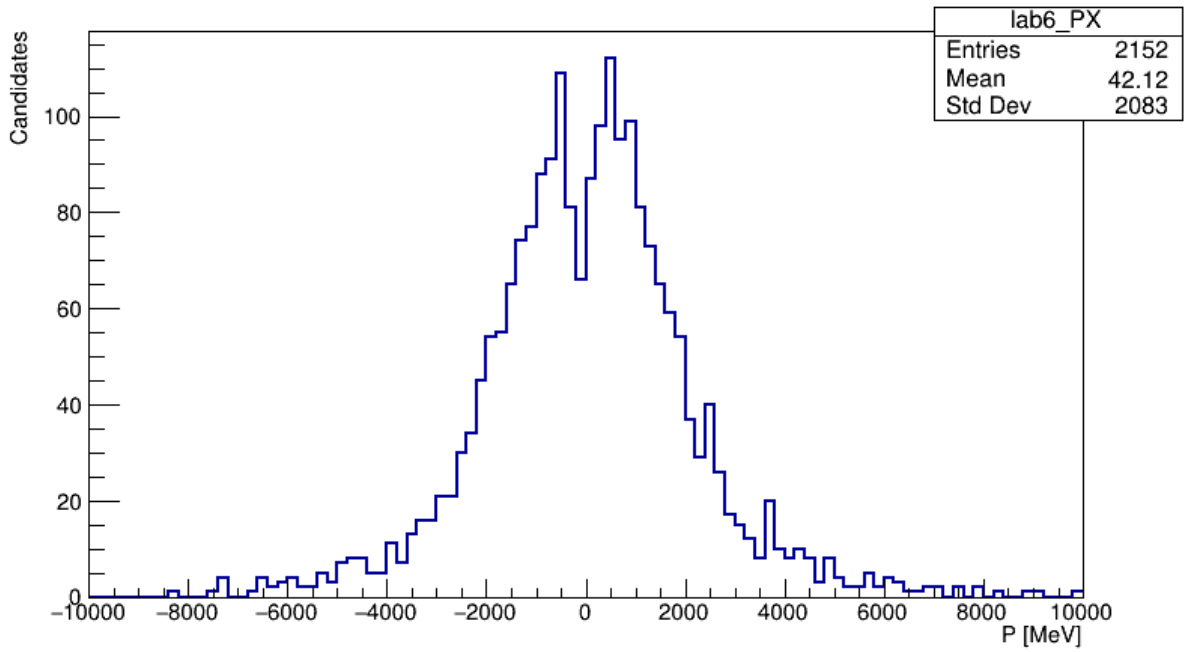
Oprócz wyżej wymienionych charakterystyk, analizie poddano również różnorakie rozkłady sygnału oraz tła wygenerowane przez obydwa klasyfikatory, celem zaobserwowania poprawy dokonanych podziałów w programie wzmocnionym danymi powielonymi. Szczegółowe omówienie analizowanych charakterystyk zawarte zostało w rozdziale 6.5.

6. Omówienie wyników

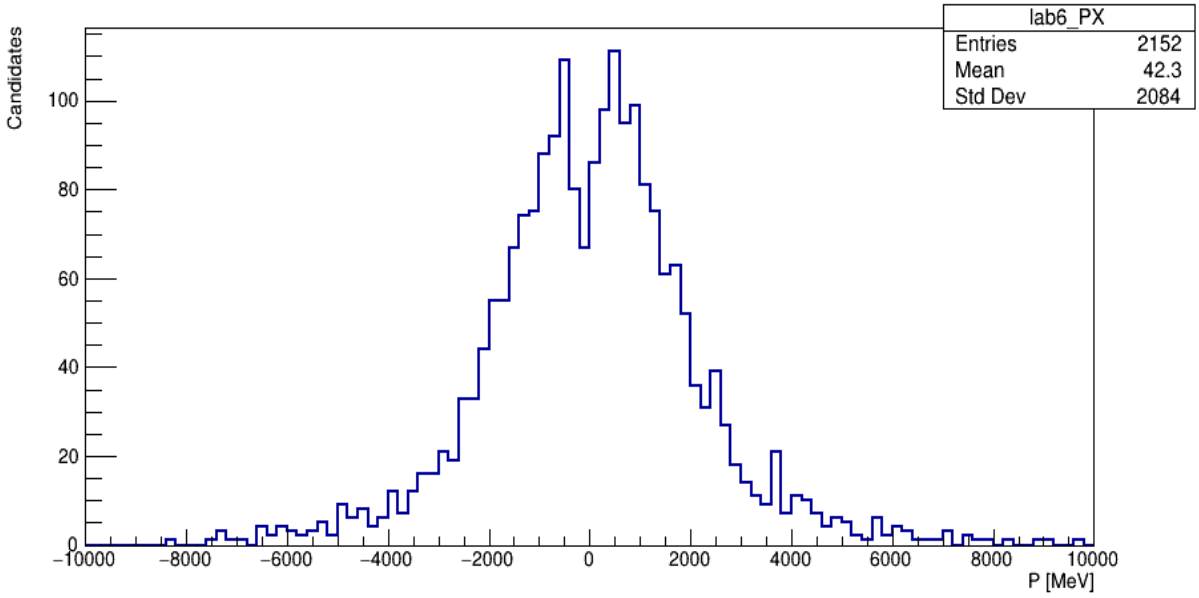
Do przedstawienia wyników multiplikacji przypadków rozpadu $B_s^0 \to D_s^* K^*$ wykorzystano histogramy rozkładu charakterystycznych atrybutów nowych rekordów, uzyskane w wyniku działania programu powielającego. Rozdział ten został podzielony na cztery części, zgodnie z czterema gałęziami algorytmu powielającego, przedstawionymi w rozdziale 5.2.

6.1. Multiplikacja - pędy stanów końcowych i rekonstrukcja drzewa rozpadu

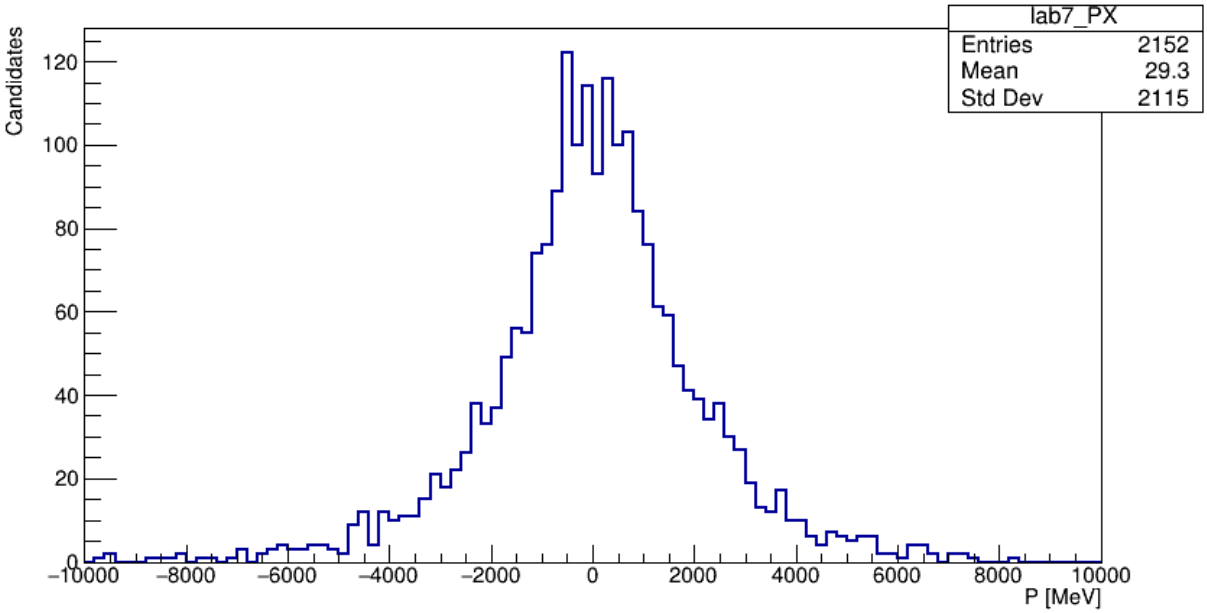
Na rysunkach 12 - 17 przedstawiono niektóre rozkłady pędów (wzdłuż osi X) cząstek $K^{\pm}$ / $\pi^{\pm}$ powstałych w wyniku rozpadu mezonu D_S dla pierwotnego, nie powielonego zbioru przypadków (nazywane dalej **pędami stanów końcowych**). Rysunki te można podzielić na dwie grupy - rysunki 12, 14, 16 obrazują rozkład pędów stanów końcowych bezpośrednio po ich wygenerowaniu przez działający przy LHCb generator przypadków, zaś rysunki 13, 15, 17 przedstawiają ten sam zestaw przypadków po przejściu przez etap symulacji efektów aparaturowych. Wybrany do prezentacji wyników zbiór danych liczy 2152 przypadki rozpadu. Pęd i energia wyrażone są w jednostkach naturalnych (MeV). Dla wszystkich trzech stanów końcowych można zauważyć, że ukazane rozkłady przed i po symulacji efektów aparaturowych są do siebie bardzo zbliżone, zarówno kształtem histogramu jak i wartościami statystyk (średnia oraz odchylenie standardowe).



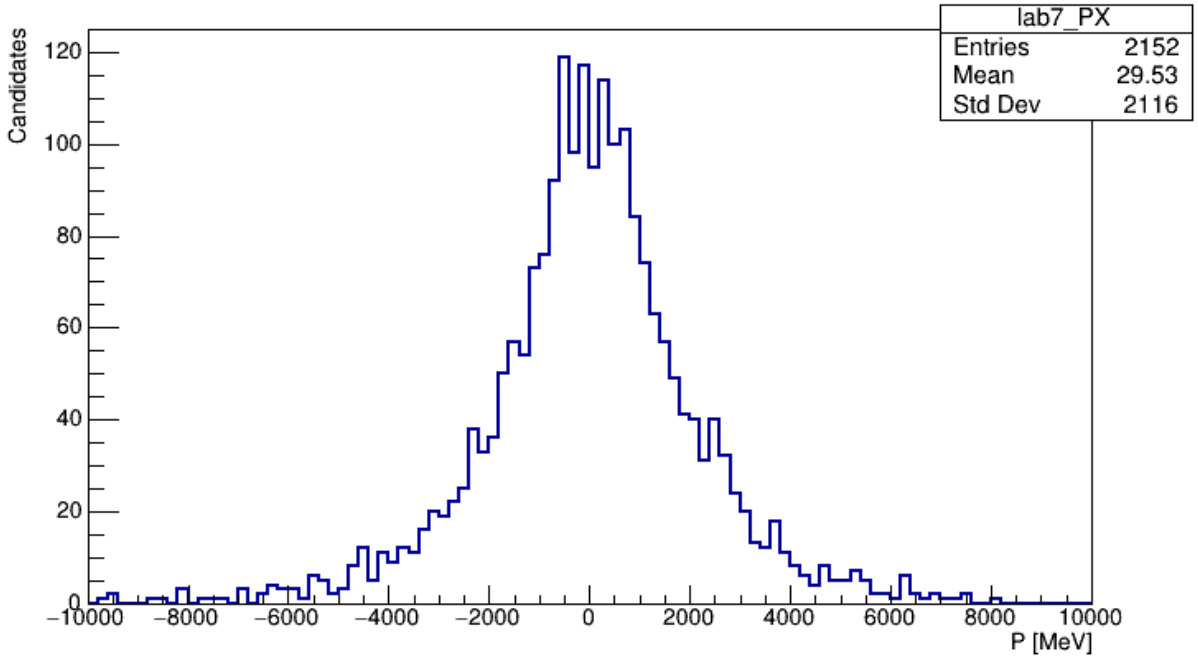
Rysunek 12 - Rozkład pędów względem osi X przed symulowaną modulacją efektów aparaturowych dla pierwszego stanu końcowego powstałego w wyniku rozpadu mezonu Ds



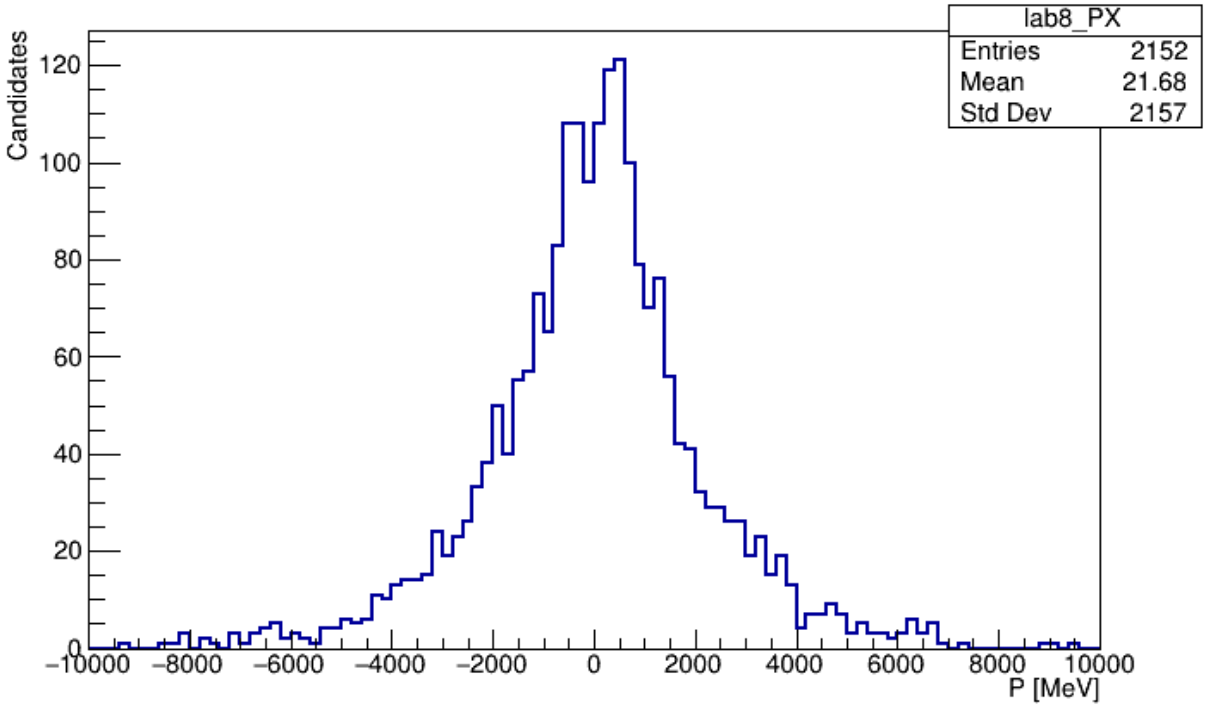
Rysunek 13 - Rozkład pędów względem osi X po symulowanej modulacji efektów aparaturowych dla pierwszego stanu końcowego powstałego w wyniku rozpadu mezonu Ds



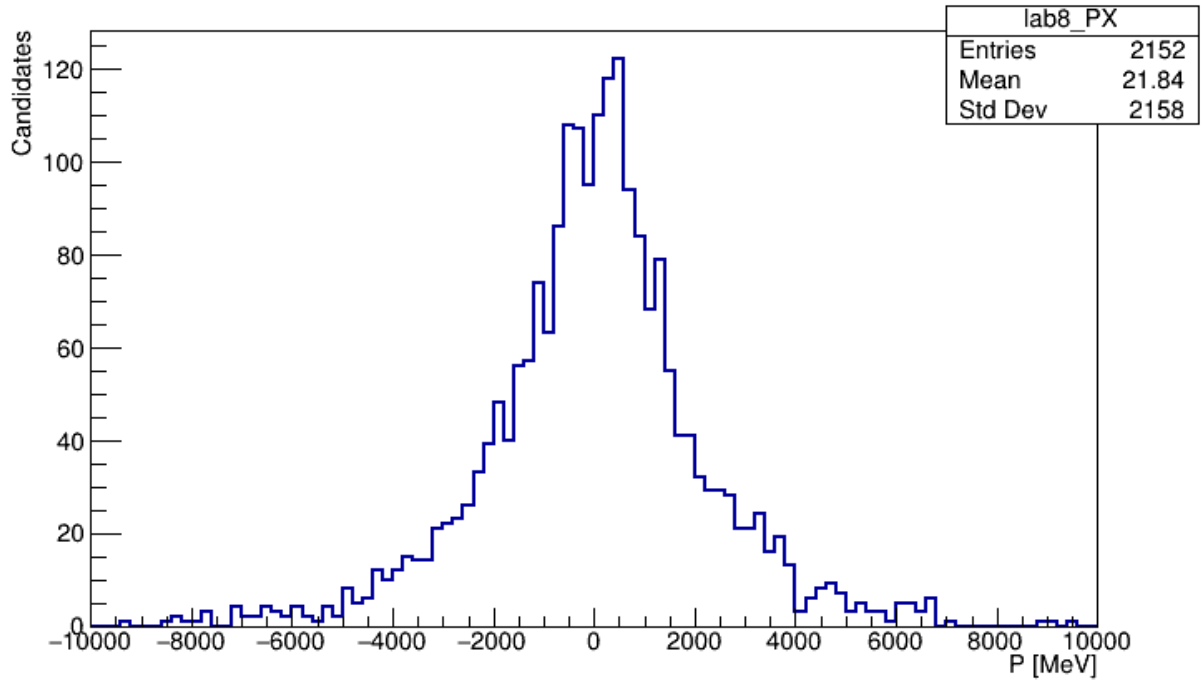
Rysunek 14 - Rozkład pędów względem osi X przed symulowaną modulacją efektów aparaturowych dla drugiego stanu końcowego powstałego w wyniku rozpadu mezonu Ds



Rysunek 15 - Rozkład pędów względem osi X po symulowanej modulacji efektów aparaturowych dla drugiego stanu końcowego powstałego w wyniku rozpadu mezonu Ds



Rysunek 16 - Rozkład pędów względem osi X przed symulowaną modulacją efektów aparaturowych dla trzeciego stanu końcowego powstałego w wyniku rozpadu mezonu Ds



Rysunek 17 - Rozkład pędów względem osi X po symulowanej modulacji efektów aparaturowych dla trzeciego stanu końcowego powstałego w wyniku rozpadu mezonu Ds

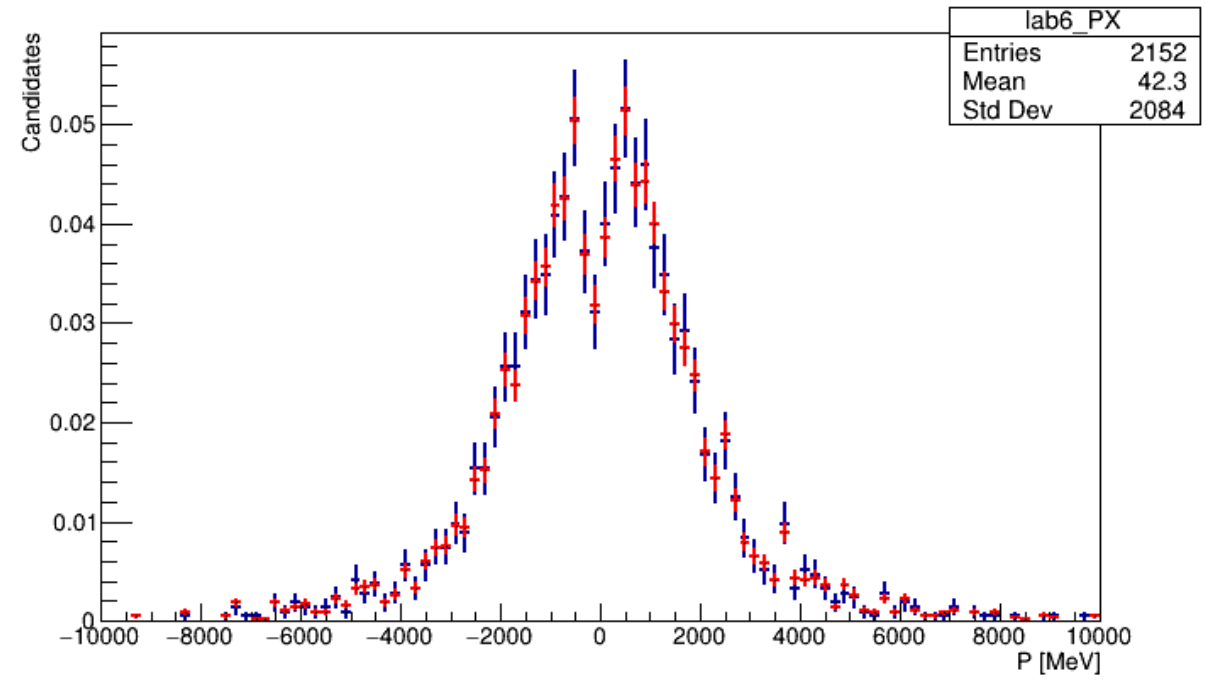
Powyższe rozkłady pokazują, że wpływ wprowadzonego rozmycia detektorowego na zarejestrowane wartości jest niewielki - dla wszystkich trzech

przedstawionych atrybutów, histogramy przed i po powieleniu są statystycznie podobne. Fakt ten jest zgodny z oczekiwaniami - gdyby było inaczej znaczyłoby to, że rejestrowane dane ulegają dużemu zniekształceniu podczas detekcji, a to prowadziłoby do przekłamań i błędnych założeń, zaś wiarygodność wniosków uzyskanych w wyniku obróbki takich danych byłaby łatwa do podważenia. Powyższa obserwacja stoi u podstaw działania algorytmu multiplikującego, i jest - pośrednio lub bezpośrednio - wykorzystywana we wszystkich czterech jego gałęziach.

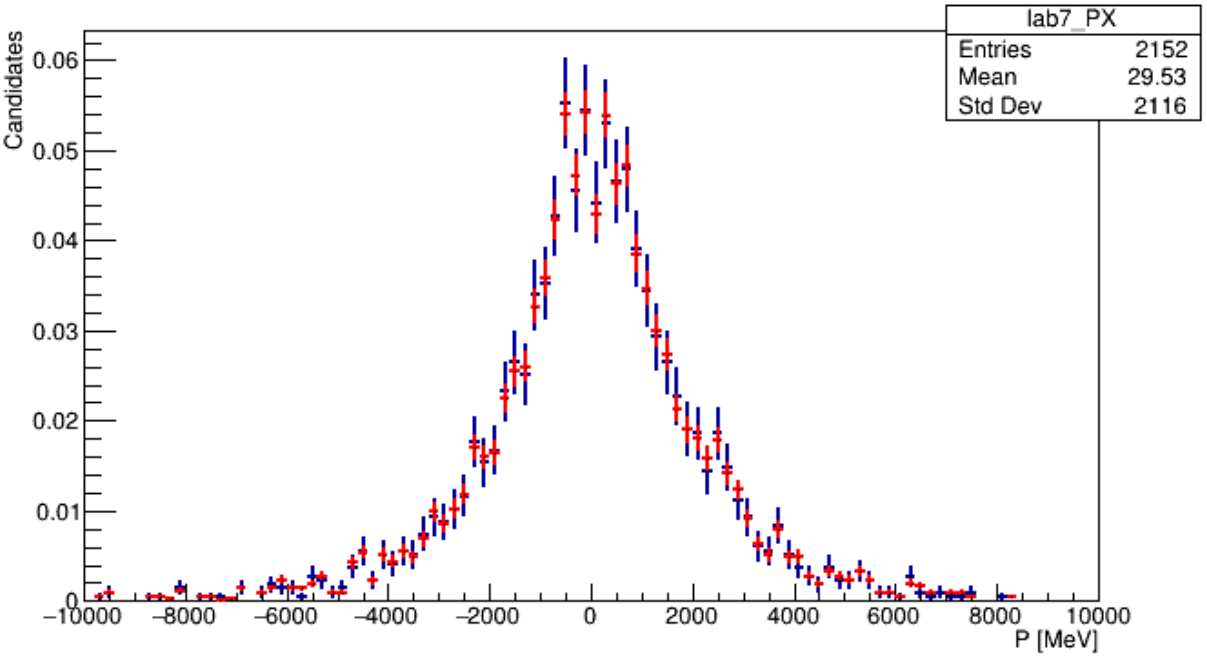
Dysponując potwierdzeniem przyjętych założeń, możliwe staje się obliczenie residuum, dającego liczbową informację o sposobie rejestracji cząstek przez detektor, jak i jego wpływie na wartości rejestrowanych atrybutów. Rozkłady residuum uzyskane dla omawianych pędów stanów końcowych, wraz z krzywymi dopasowania rozkładu prawdopodobieństwa przedstawione zostały wcześniej, na rysunku 7.

Planując obliczenie wpływu etapu symulowanej detekcji na rozkład wartości badanych cech cząstek, oczekiwano że wpływ ten będzie niewielki. Powyższe histogramy potwierdzają tę tezę. Otrzymane wartości residuum oscylują wokół zera, a ich rozrzut (mierzony wielkością odchylenia standardowego) jest 200-krotnie mniejszy od odchylenia standardowego danych użytych do obliczeń. Krzywe dopasowania rozkładu prawdopodobieństwa, wyznaczone dla otrzymanych histogramów wskazują na rozkład normalny.

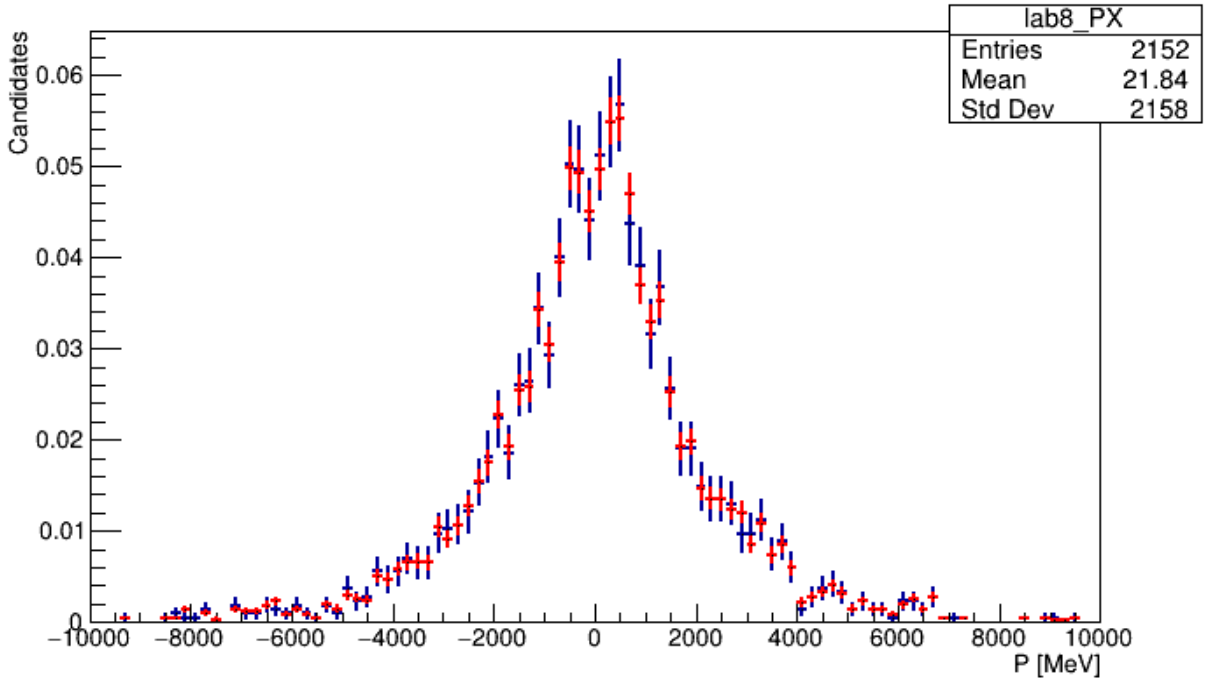
Wykorzystując otrzymane statystyki pędów stanów końcowych, przystąpiono do następnego kroku algorytmu - multiplikacji i rekonstrukcji drzewa rozpadu na podstawie powielonych wartości. Celem potwierdzenia poprawności zaimplementowanego algorytmu multiplikującego, na histogramy zawierające dane wejściowe nałożone zostały rozkłady uzyskane w wyniku czterokrotnego powielenia, a całość poddano normalizacji, aby umożliwić bezpośrednie porównanie rozkładów na jednym histogramie. Wyniki tych działań przedstawiają rysunki 18 - 20



Rysunek 18 - Porównanie znormalizowanych rozkładów pędów stanów końcowych wzdłuż osi X pochodzących z rozpadu mezonu Ds dla pierwszego stanu końcowego - punkty niebieskie: próbka pierwotna, punkty czerwone: próbka powielona



Rysunek 19 - Porównanie znormalizowanych rozkładów pędów stanów końcowych wzdłuż osi X pochodzących z rozpadu mezonu Ds dla drugiego stanu końcowego - punkty niebieskie: próbka pierwotna, punkty czerwone: próbka powielona

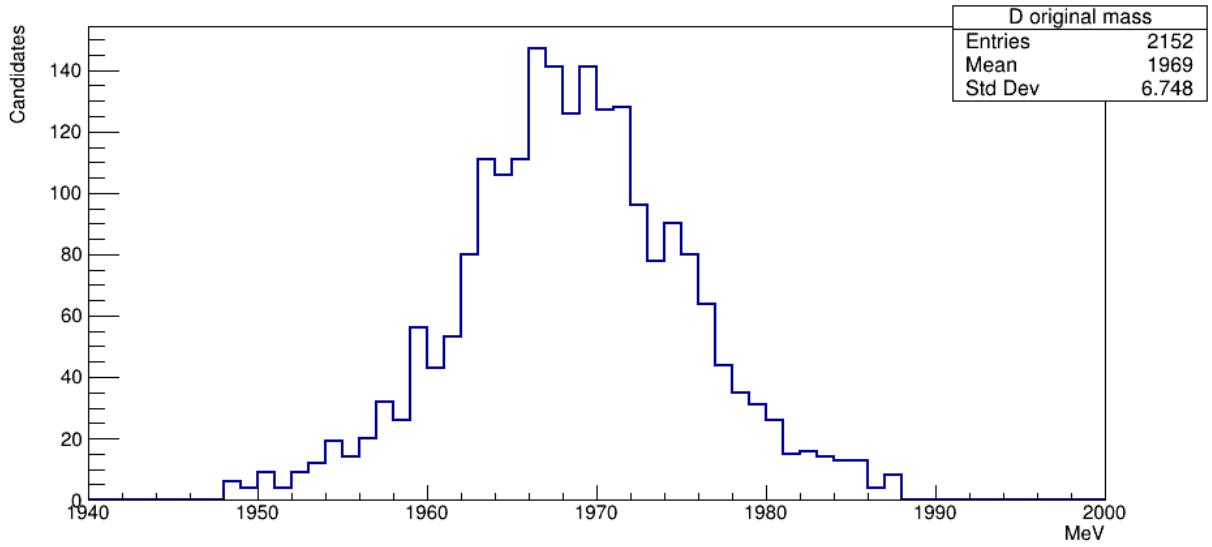


Rysunek 20 - Porównanie znormalizowanych rozkładów pędów stanów końcowych wzdłuż osi X pochodzących z rozpadu mezonu Ds dla trzeciego stanu końcowego - punkty niebieskie: próbka pierwotna, punkty czerwone: próbka powielona

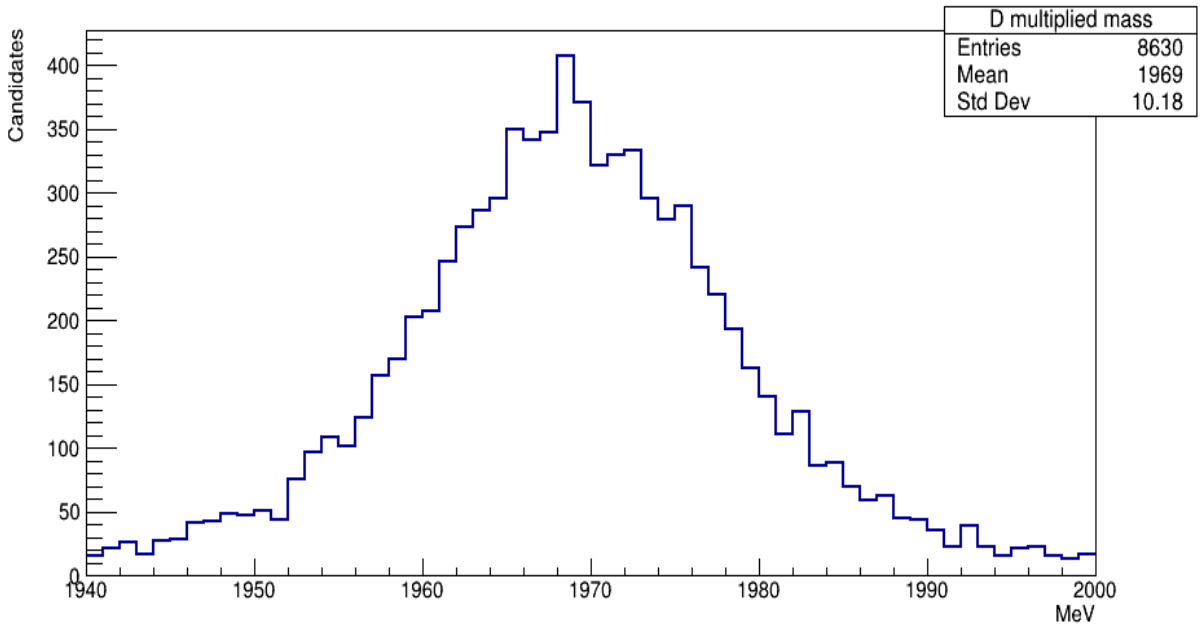
Przedstawione wyniki wskazują na zgodność pomiędzy próbkami źródłowymi i powielonymi, co wskazuje na poprawne działanie algorytmu. Można zauważyć, że próbka powielona jest niemal identyczna z próbką pierwotną, co doskonale wpasowuje się w ramy przyjętego w algorytmie założenia o "nieznacznej modyfikacji" danych wejściowych.

Zgodnie z przedstawionym wcześniej algorytmem, następnym krokiem była rekonstrukcja drzewa rozpadu na podstawie nowo uzyskanych przypadków. Głównym celem rekonstrukcji jest rozszerzenie spektrum działana algorytmu multiplikującego na cząstki, których parametry nie są *explicite* poddane całemu żmudnemu, opisanemu powyżej procesowi multiplikacji. Dysponując podstawową wiedzą fizyczną dotyczącą pędu, energii oraz zderzeń, można "odtworzyć" przebieg rozpadów $B_s^0 \rightarrow D_s^* K^*$ których końcowym efektem byłyby stany końcowe o takich pędach, jak te uzyskane w wyniku powielenia. Dodatkowym, jednakże bardzo pożądanym, efektem rekonstrukcji drzewa rozpadu jest możliwość sprawdzenia poprawności przeprowadzonych multiplikacji poprzez analizę histogramów mas poszczególnych

cząstek powstających w zderzeniu (D_S , D_S^* , oraz B_s) dla oryginalnego zbioru danych, oraz dla próbki czterokrotnie powielonej. Histogramy te zostały przedstawione na rysunkach 21 - 26.

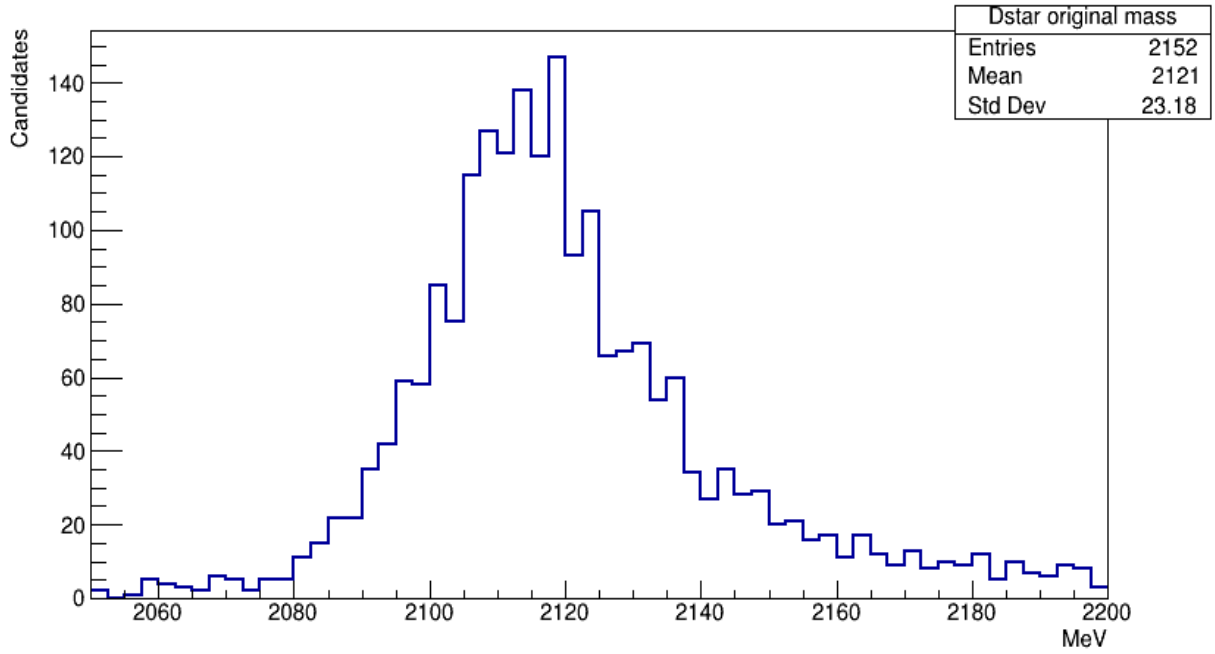


Rysunek 21 - Histogram masy mezonu Ds przed powieleniem metodą rekonstrukcji drzewa rozpadu

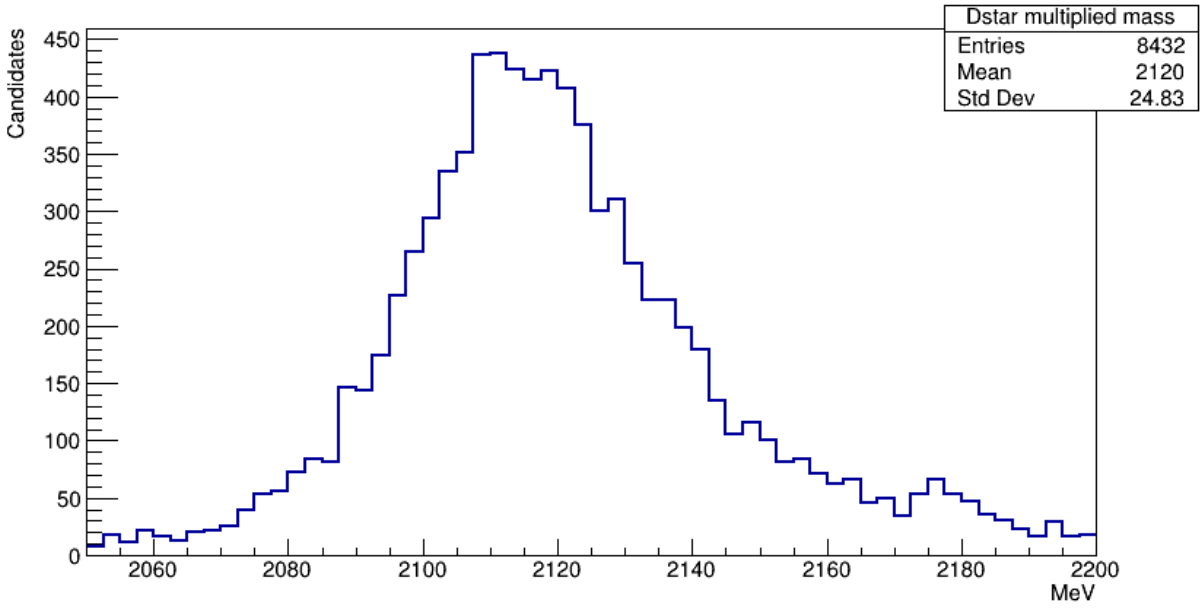


Rysunek 22 - Histogram masy mezonu Ds po powieleniu metodą rekonstrukcji drzewa rozpadu

Rozkłady masy mezonu D_S przed i po powieleniu w wyniku rekonstrukcji drzewa rozpadu są do siebie zbliżone. Ich wartość średnia jest identyczna, i zgodna z wartością tablicową 1968.8 MeV [12], natomiast odchylenie standardowe próbki powielonej jest o 51% większe, co skutkuje zauważalnym spłaszczeniem i rozciągnięciem odpowiadającego jej histogramu. Prawdopodobną przyczyną takiego stanu rzeczy są nieznaczne różnice w przedstawionych powyżej rozkładach pędów cząstek - córek. Niepewność wynikająca z rozmycia parametrów trzech cząstek sumarycznie skutkuje zwiększeniem obserwowanej niepewności wartości masy mezonu D_S. Otrzymany rozkład wykorzystano do dalszej rekonstrukcji drzewa rozpadu, a w efekcie uzyskania histogramu masy stanu rezonansowego D_s^* dla powielonych przypadków rozpadu, i porównano go z rozkładem wynikającym z próbki pierwotnej (rysunki 23, 24). Większy rozrzut wartości wynikający ze zwiększonego odchylenia standardowego masy mezonu D_S nie ma odzwierciedlenia w rozkładzie powielonym dla stanu rezonansowego D_s^*.

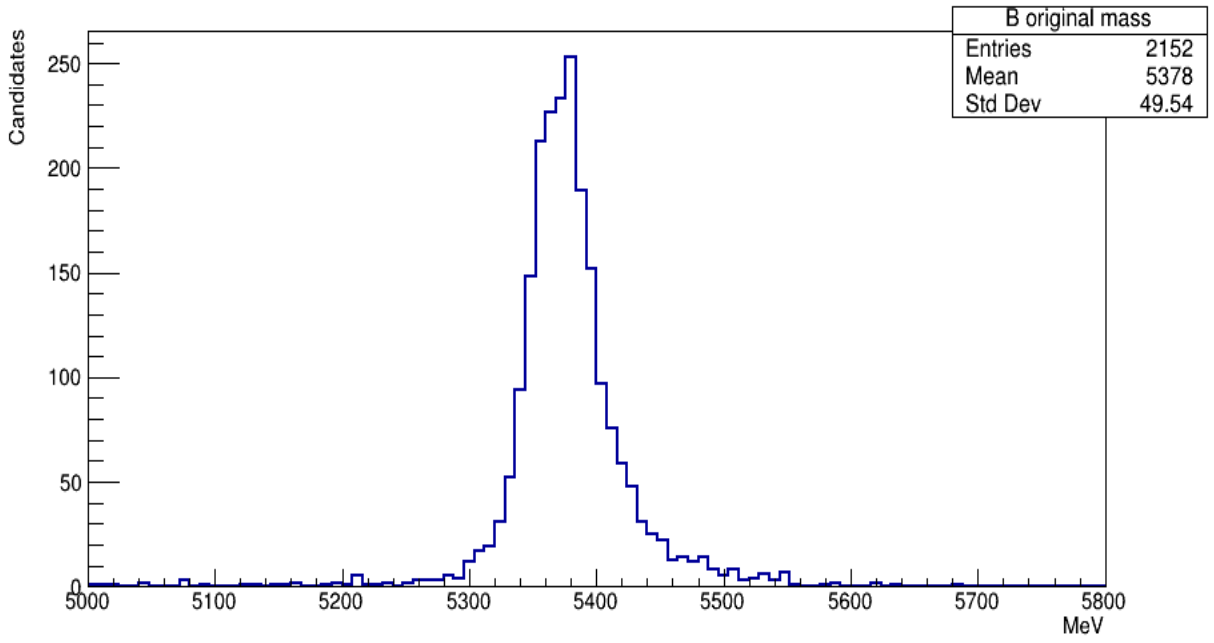


Rysunek 23 - Histogram masy stanu rezonansowego D_S^* przed powieleniem metodą rekonstrukcji drzewa rozpadu

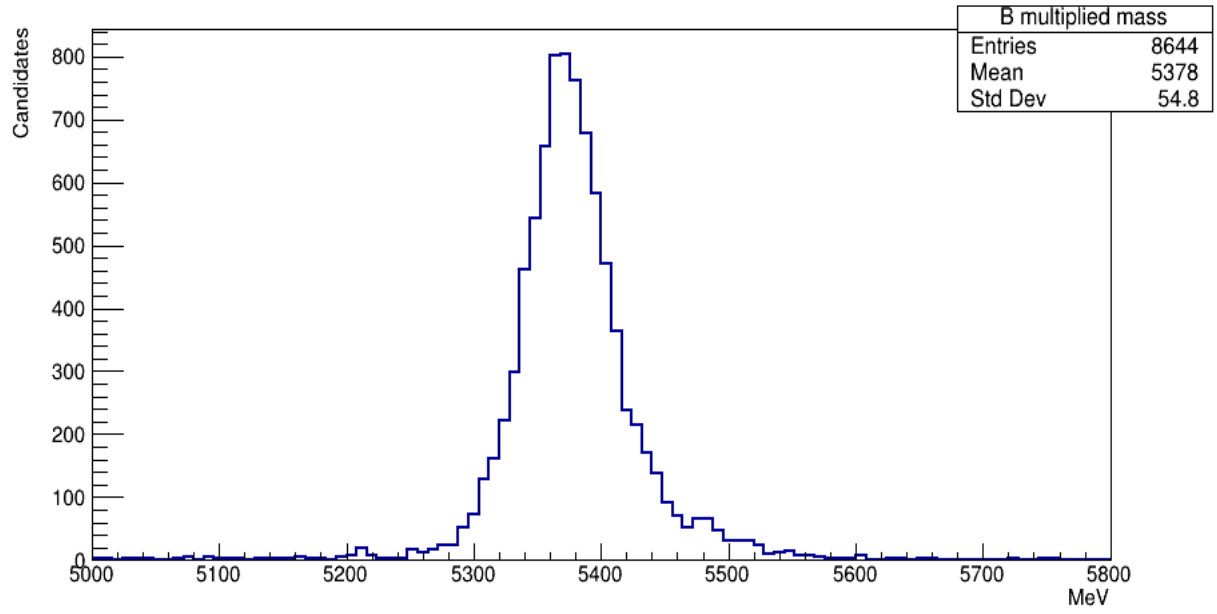


Rysunek 24 - Histogram masy stanu rezonansowego D_S^* po powieleniu metodą rekonstrukcji drzewa rozpadu

Choć powyższe histogramy różnią się nieco kształtem, ich statystyki - wartość średnia i odchylenie standardowe są niemal identyczne. Wartości główne obu histogramów (odpowiednio 2121 oraz 2120 MeV) są blisko wartości tablicowej, wynoszącej 2112 MeV [12]. Ostatecznym potwierdzeniem prawidłowości otrzymanych wyników jest analiza rozkładów masy mezonu $\boldsymbol{B_s}$ (rysunki 25, 26).



Rysunek 25 - Histogram masy mezonu B_S przed powieleniem metodą rekonstrukcji drzewa rozpadu



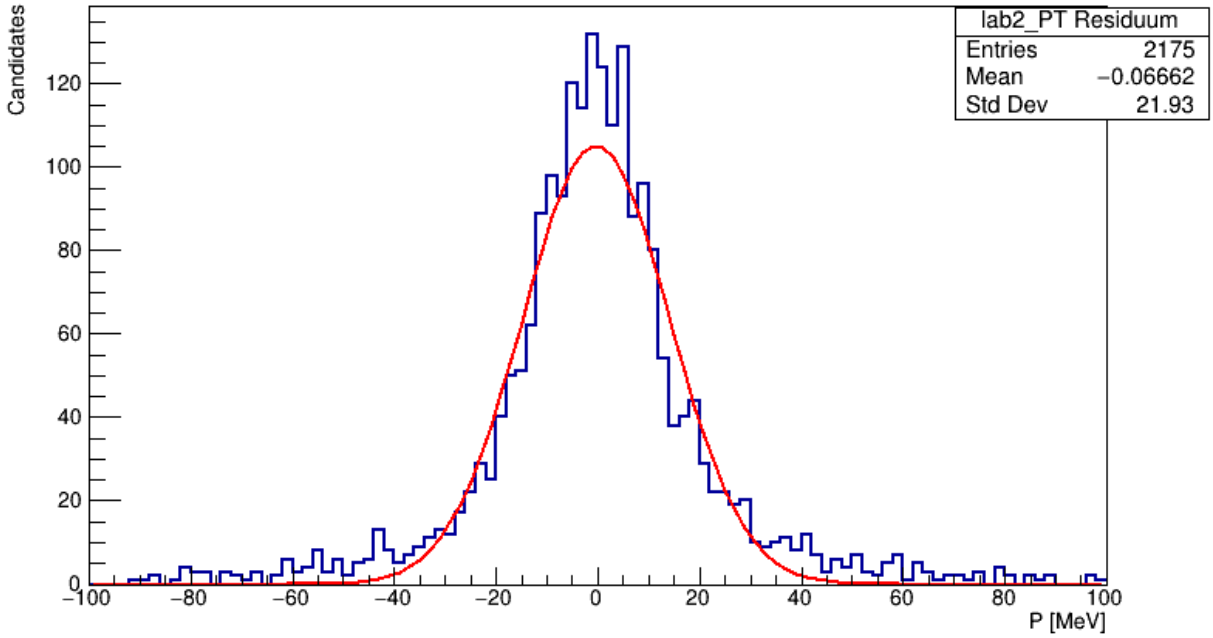
Rysunek 26 - Histogram masy mezonu B_S po powieleniu metodą rekonstrukcji drzewa rozpadu

Statystyczne parametry rozkładu masy mezonu $\boldsymbol{B_S}$ przed i po multiplikacji są do siebie bardzo zbliżone. Wartość średnia, wynosząca w obu przypadkach 5378 MeV jest zgodna z wartością tablicową 5366 MeV [12]. Oznacza to, że na wszystkich etapach rekonstrukcji drzewa rozpadu otrzymane wyniki wskazują na poprawność

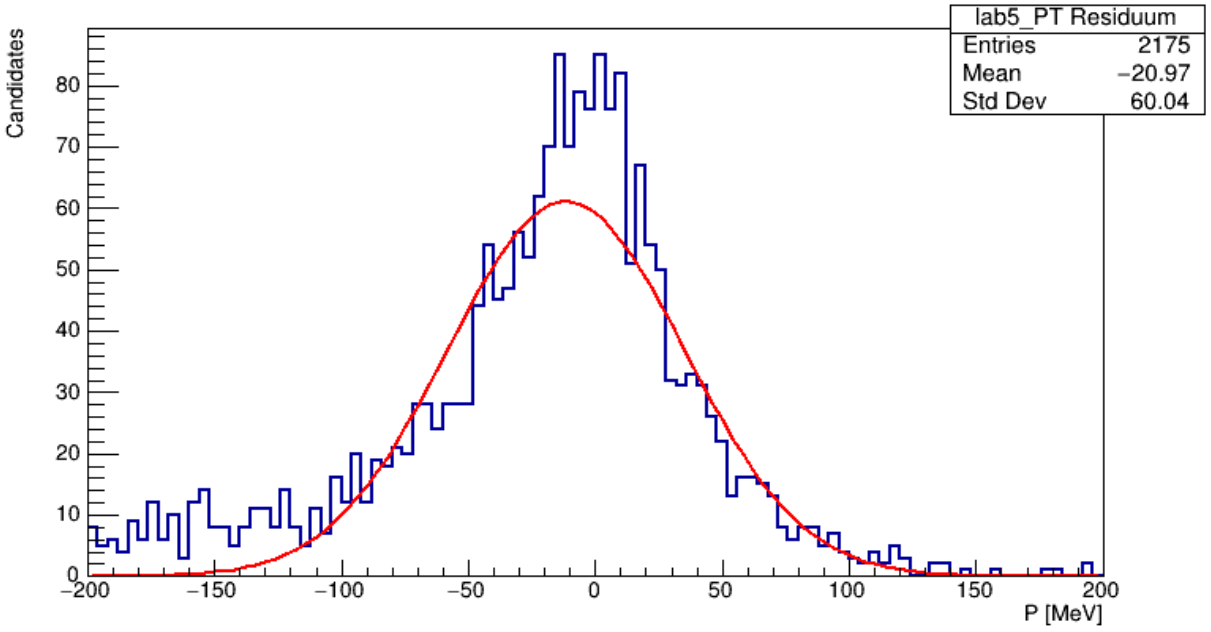
implementacji algorytmu powielającego w części dotyczącej multiplikacji poprzez modyfikację wartości pędów stanów końcowych.

6.2. Multiplikacja - pędy poprzeczne

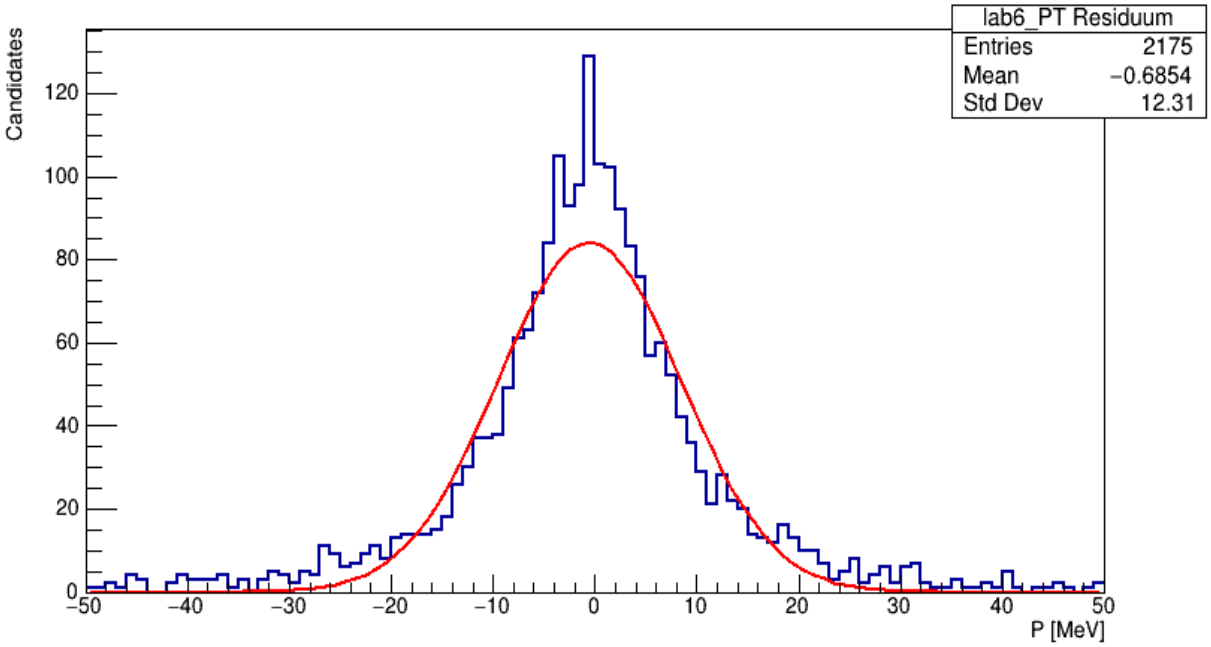
Podobnie jak w poprzednim punkcie, również w tej gałęzi algorytmu sprawdzenie fizycznej poprawności nowo uzyskanych przypadków odbyło się na zasadzie porównania wartości pędów poprzecznych dla poszczególnych cząstek otrzymanych z próbki pierwotnej, z wartościami uzyskanymi w wyniku działania algorytmu powielającego. Krokiem pośrednim niezbędnym do powielenia rekordów było wyliczenie wartości residuum dla każdej z procesowanych cząstek. Rysunki 27 - 29 przedstawiają jego rozkłady dla niektórych cząstek, wykorzystane do stworzenia generatora odpowiedzi detektora.



Rysunek 27 - Rozkład wartości residuum obliczonych dla pędu poprzecznego cząstki K_S^0 wraz z dopasowanym rozkładem normalnym



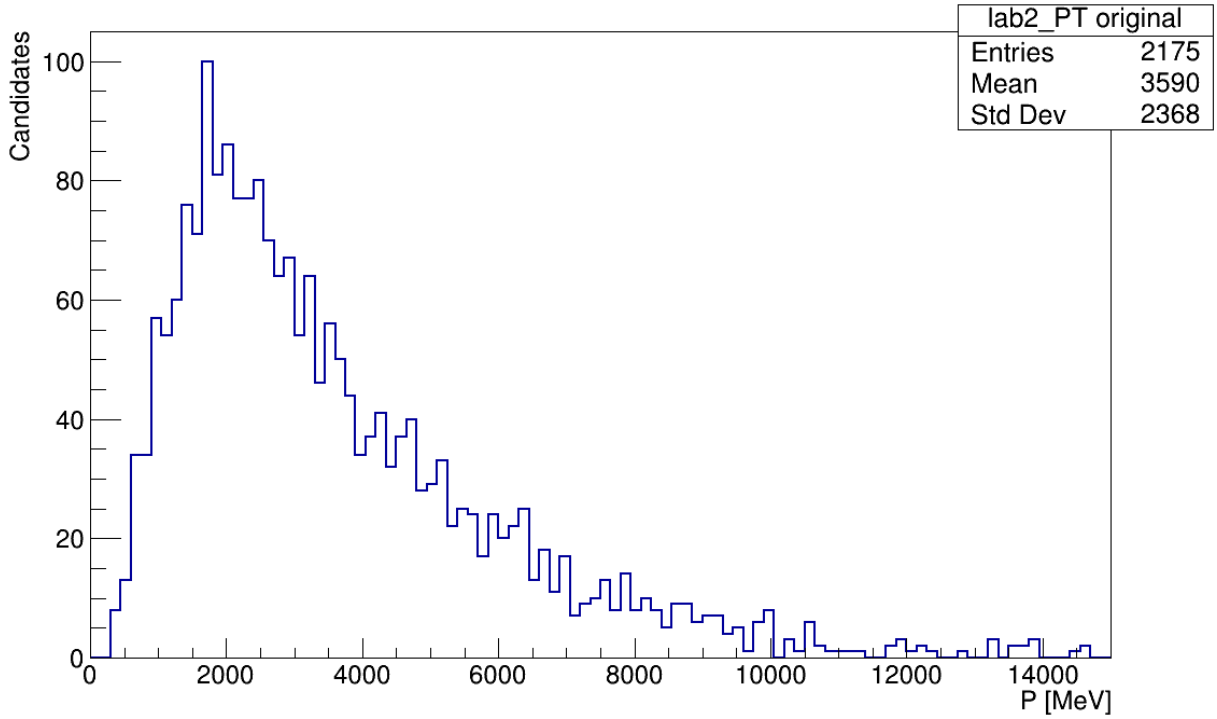
Rysunek 28 - Rozkład wartości residuum obliczonych dla pędu poprzecznego cząstki γ wraz z dopasowanym rozkładem normalnym



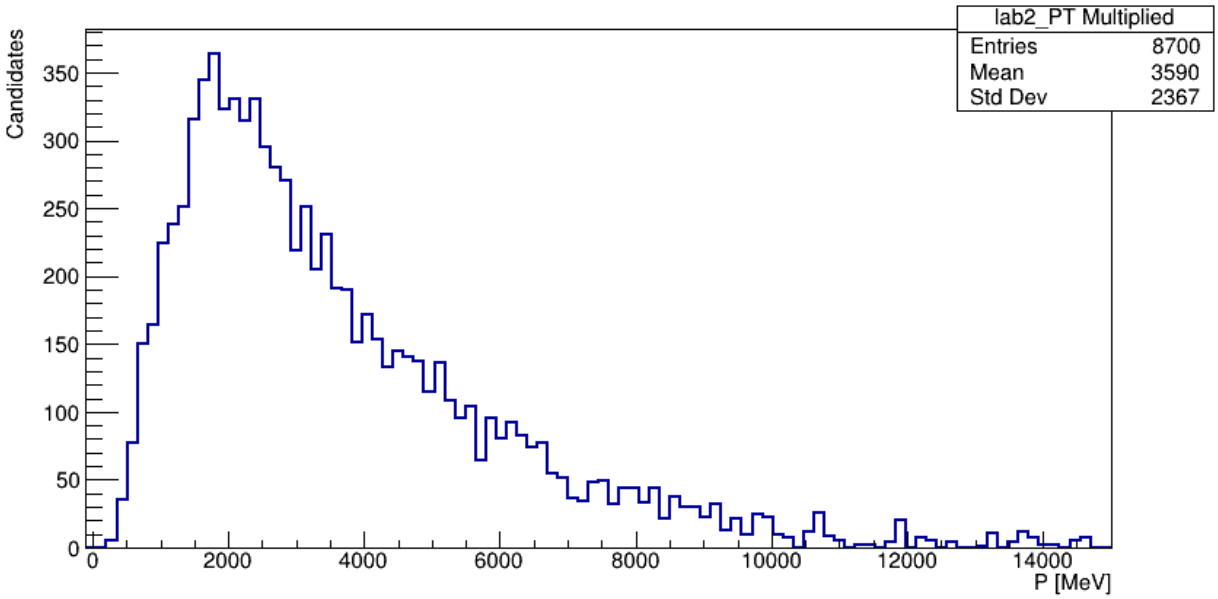
Rysunek 29 - Rozkład wartości residuum obliczonych dla pędu poprzecznego pierwszego mezonu ($K^{\pm}$ / $\pi^{\pm}$) pochodzącego z rozpadu mezonu Ds wraz z dopasowanym rozkładem normalnym

Powyższe przykładowe rozkłady residuum potwierdzają obserwacje przedstawione w rozdziale 6.1. Dla wszystkich powielanych pędów poprzecznych z wyjątkiem cząstki γ, wartość średnia residuum zawiera się w przedziale (-1.0 ; 1.0), a odchylenie standardowe nie przekracza 70 MeV. Otrzymane wartości residuum są

bardzo małe w porównaniu z oryginalnymi wartościami zmiennych dla których zostały wyliczone - w przypadku każdej zmiennej mówimy o co najmniej dwóch - trzech rzędach wielkości (przykładowo - dla pędu poprzecznego pierwszego mezonu ($K^{\pm}$ / $\pi^{\pm}$) pochodzącego z rozpadu mezonu $\boldsymbol{D_S}$, wartość średnia wynosi 2401, zaś odchylenie standardowe 1822 MeV, wobec wartości średniej residuum równej -0.6854 i odchyleniu standardowym 12.31 MeV). W każdym przypadku wartości residuum rozkładają się zgodnie z rozkładem normalnym, co pozwoliło na stworzenie statystycznego generatora modulacji efektów aparaturowych, zgodnie z zaprezentowanym algorytmem. Porównanie histogramów powielonych cech z ich wartościami występującymi w próbce pierwotnej przedstawiają rysunki 30 - 35

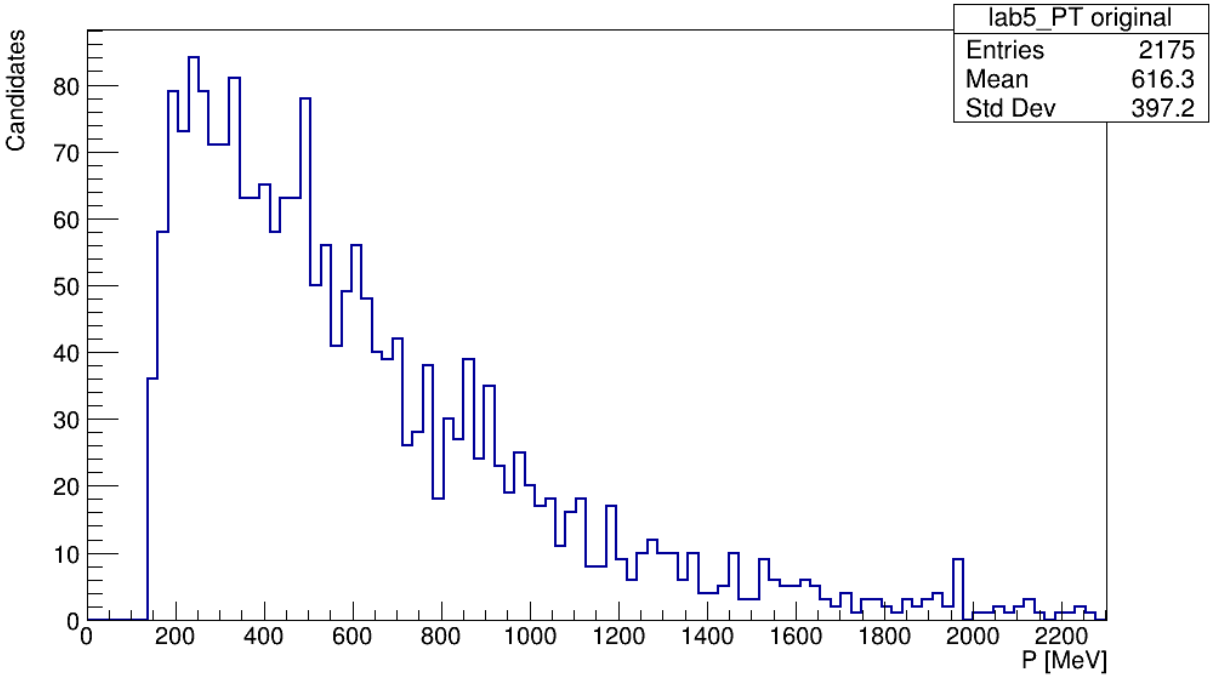


Rysunek 30 - Histogram pędu poprzecznego cząstki K_S^0 przed powieleniem metodą modulacji efektów aparaturowych

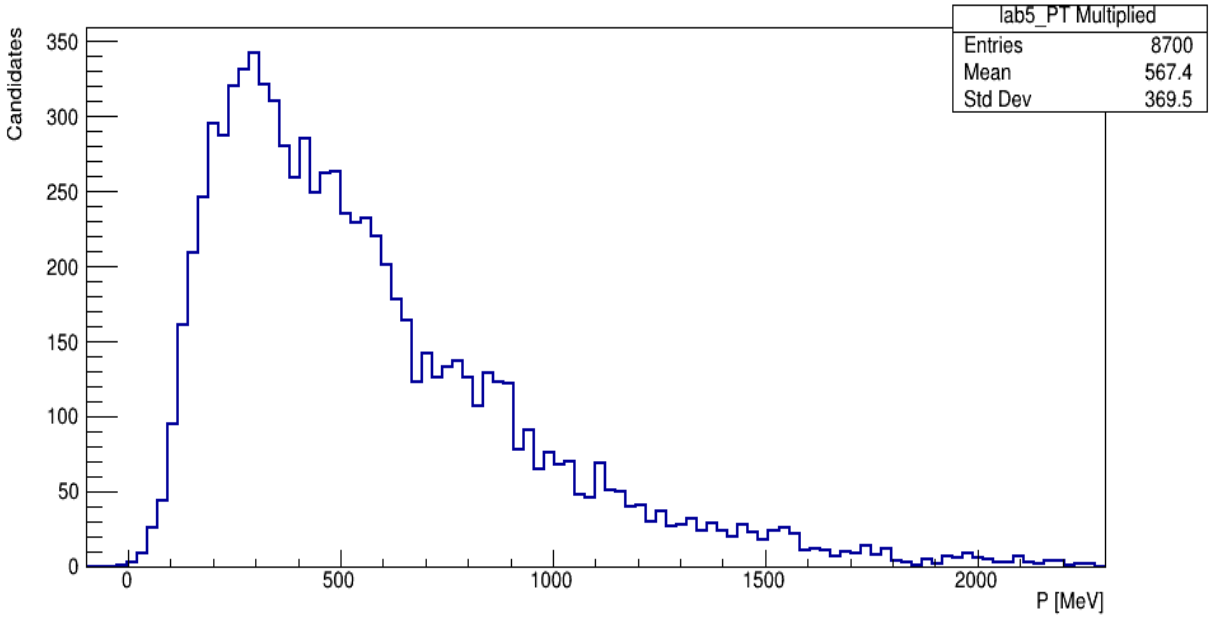


Rysunek 31 - Histogram pędu poprzecznego cząstki K_S^0 po powieleniu metodą modulacji efektów aparaturowych

W powyższym przykładzie rozkłady wartości pędu poprzecznego oryginalne oraz powielone są niemalże identyczne. Wartość średnia w obu przypadkach jest taka sama i wynosi 3590 MeV, zaś odchylenie standardowe różni się o 1 MeV (różnica rzędu 0,04%). Wyniki te sugerują, że technika powielenia stosowana dla pędów poprzecznych jest poprawna z fizycznego punktu widzenia.



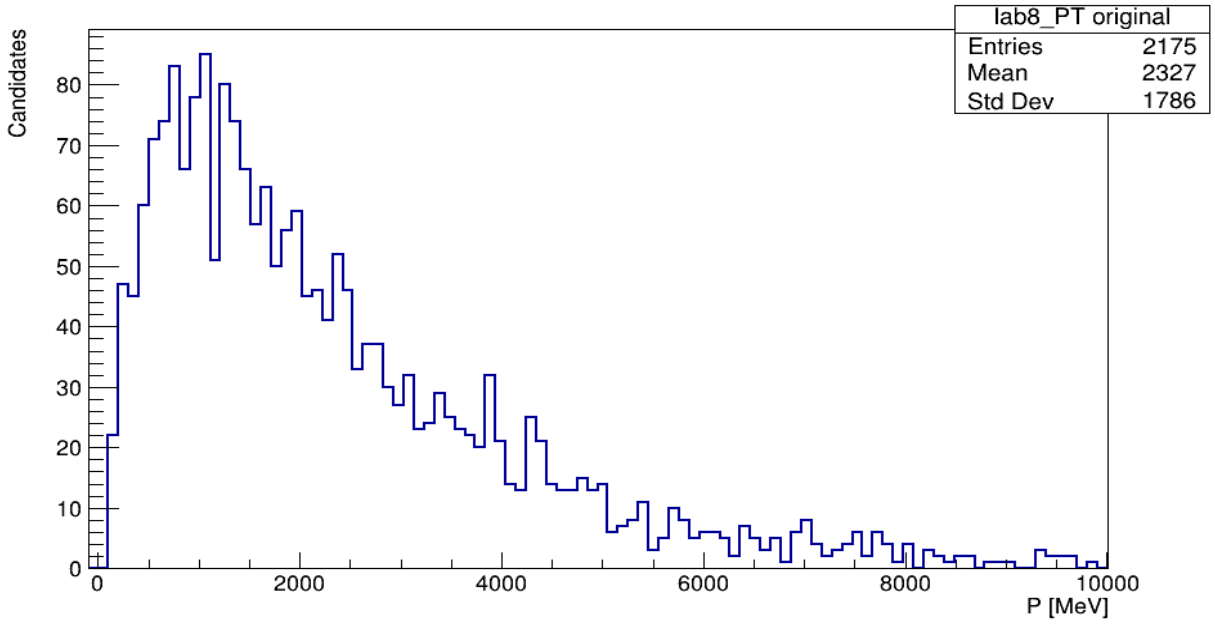
Rysunek 32 - Histogram pędu poprzecznego cząstki γ przed powieleniem metodą modulacji efektów aparaturowych



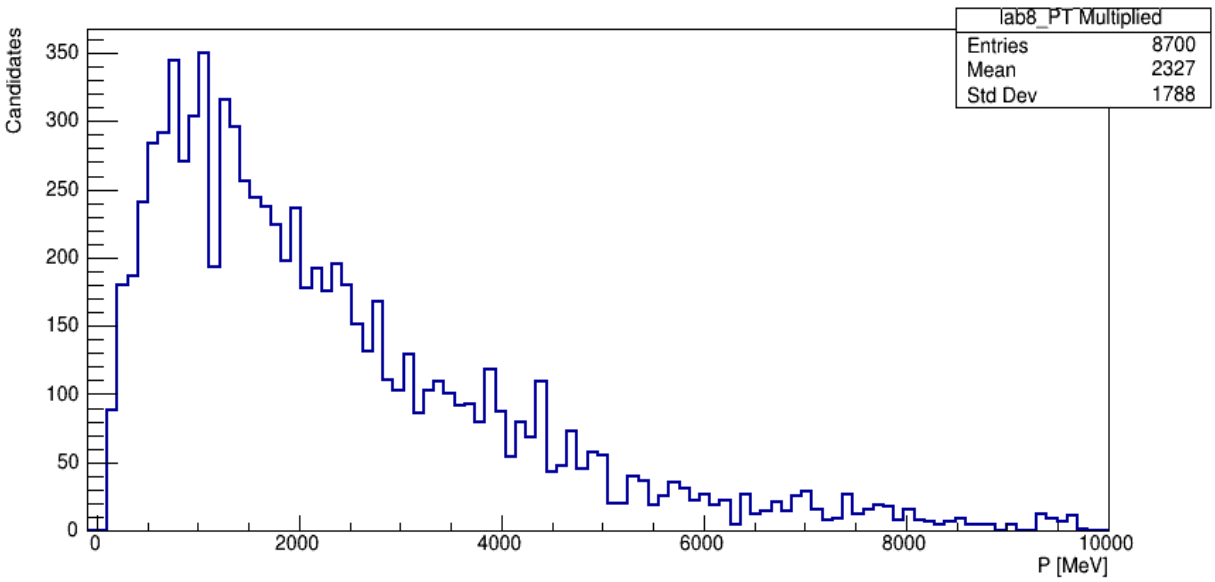
Rysunek 33 - Histogram pędu poprzecznego cząstki γ po powieleniu metodą modulacji efektów aparaturowych

W przypadku cząstki $\boldsymbol{\gamma}$ histogramy są do siebie zbliżone, jednak analiza parametrów statystycznych wskazuje, że występujące między nimi różnice są większe

niż w przypadku cząstki K_S^0. Wartość średnia próbki powielonej jest o 49 MeV mniejsza od wartości z próbki oryginalnej, zmniejszeniu uległo również odchylenie standardowe (o 28 MeV)



Rysunek 34 - Histogram pędu poprzecznego drugiego mezonu ($K^{\pm}$ / $\pi^{\pm}$) pochodzącego z rozpadu mezonu Ds przed powieleniem metodą modulacji efektów aparaturowych

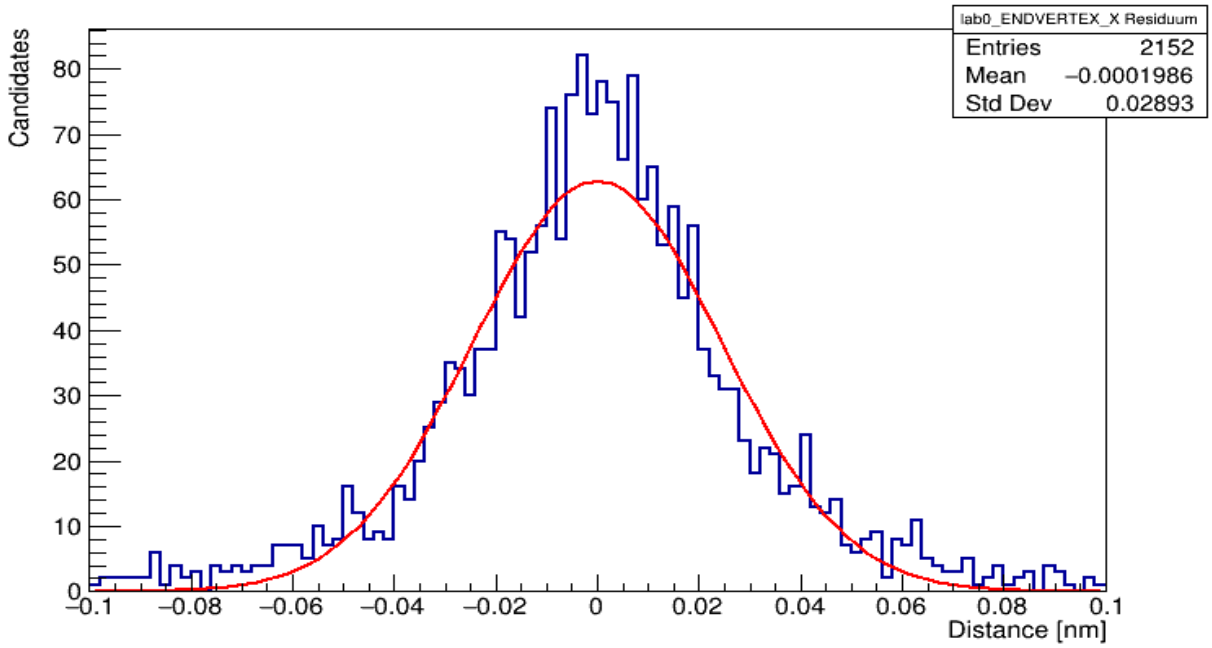


Rysunek 35 - Histogram pędu poprzecznego drugiego mezonu ($K^{\pm}$ / $\pi^{\pm}$) pochodzącego z rozpadu mezonu Ds po powieleniu metodą modulacji efektów aparaturowych

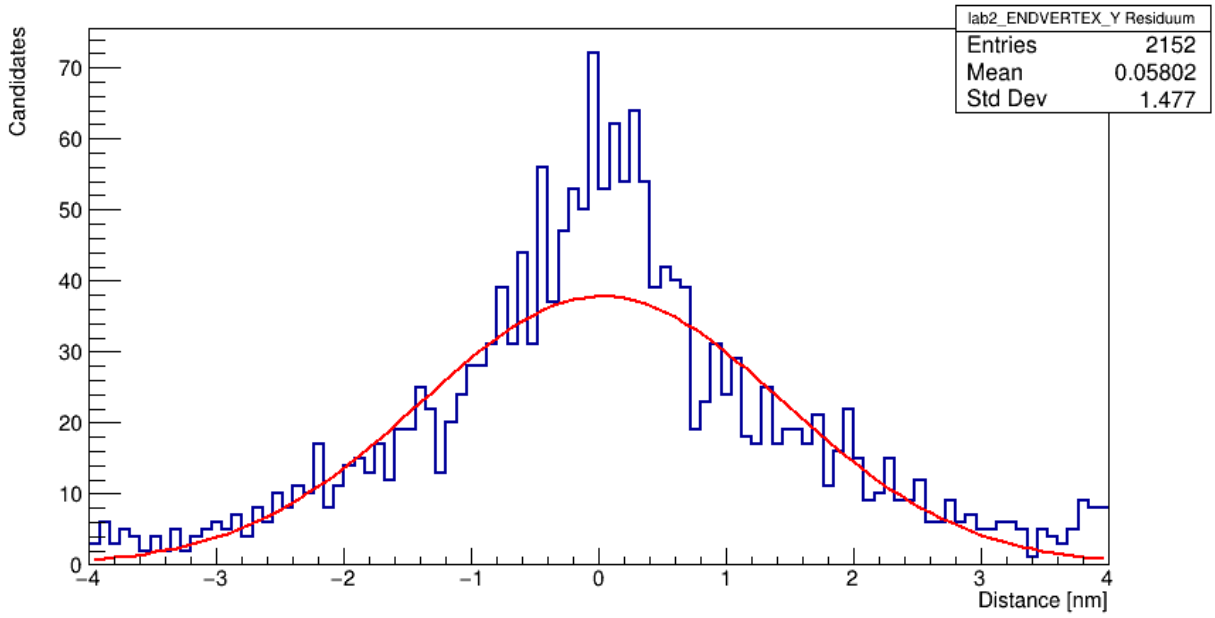
W przypadku mezonu $\boldsymbol{K^{\pm}}$ / $\boldsymbol{\pi^{\pm}}$ zarówno rozkład wartości, jak i parametry statystyczne są niemal identyczne dla obydwu histogramów. Na podstawie wszystkich przedstawionych rozkładów stwierdzono, że technika sztucznego powielenia zastosowana w przypadku pędów poprzecznych daje poprawne z fizycznego punktu widzenia wyniki, a dane uzyskane w ten sposób mogą zostać poddane weryfikacji w klasyfikatorze opartym o technikę BDT.

6.3. Współrzędne wierzchołków powstania i rozpadu cząstek

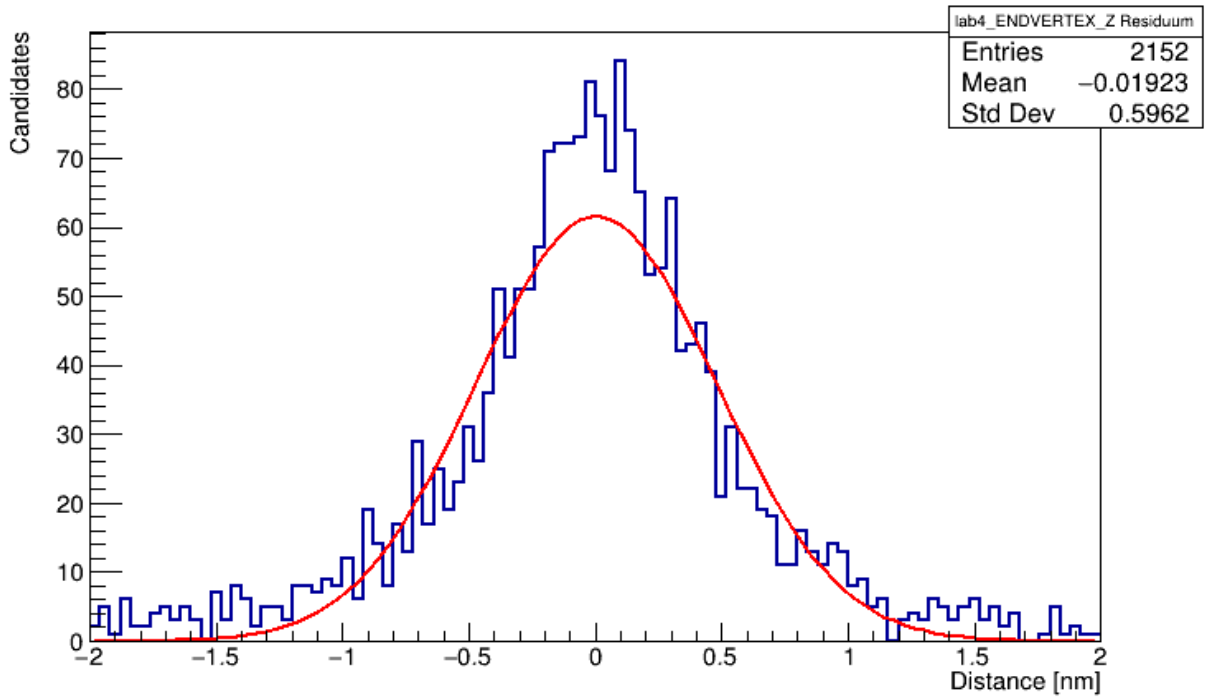
Mając w pamięci dodatkowe wymagania stawiane algorytmowi powielającemu w części dotyczącej wierzchołków powstania i rozpadu cząstek mogłoby się wydawać, że omówienie wyników tego fragmentu będzie bardziej skomplikowane. Należy jednak pamiętać, że sama logika algorytmu nie zmienia się względem wersji dla pędów poprzecznych, a dodatkowe akcje ograniczają się do nadpisania odpowiednich atrybutów cząstek wyliczonymi wartościami. Z tego względu wyniki tej części zostaną omówione w taki sam sposób jak poprzednie. Spośród wszystkich multiplikowanych wartości, w omówieniu zaprezentowane zostaną wyniki dla współrzędnej X wierzchołka rozpadu mezonu $\boldsymbol{B_S}$, współrzędnej Y wierzchołka rozpadu mezonu K_S^0 , oraz współrzędnej Z wierzchołka rozpadu mezonu $\boldsymbol{D_s}$. Wartości residuum obliczone dla wymienionych współrzędnych przedstawiają rysunki 36 - 38



Rysunek 36 - Rozkład wartości residuum obliczonych dla współrzędnej X wierzchołka rozpadu mezonu Bs wraz z dopasowanym rozkładem normalnym

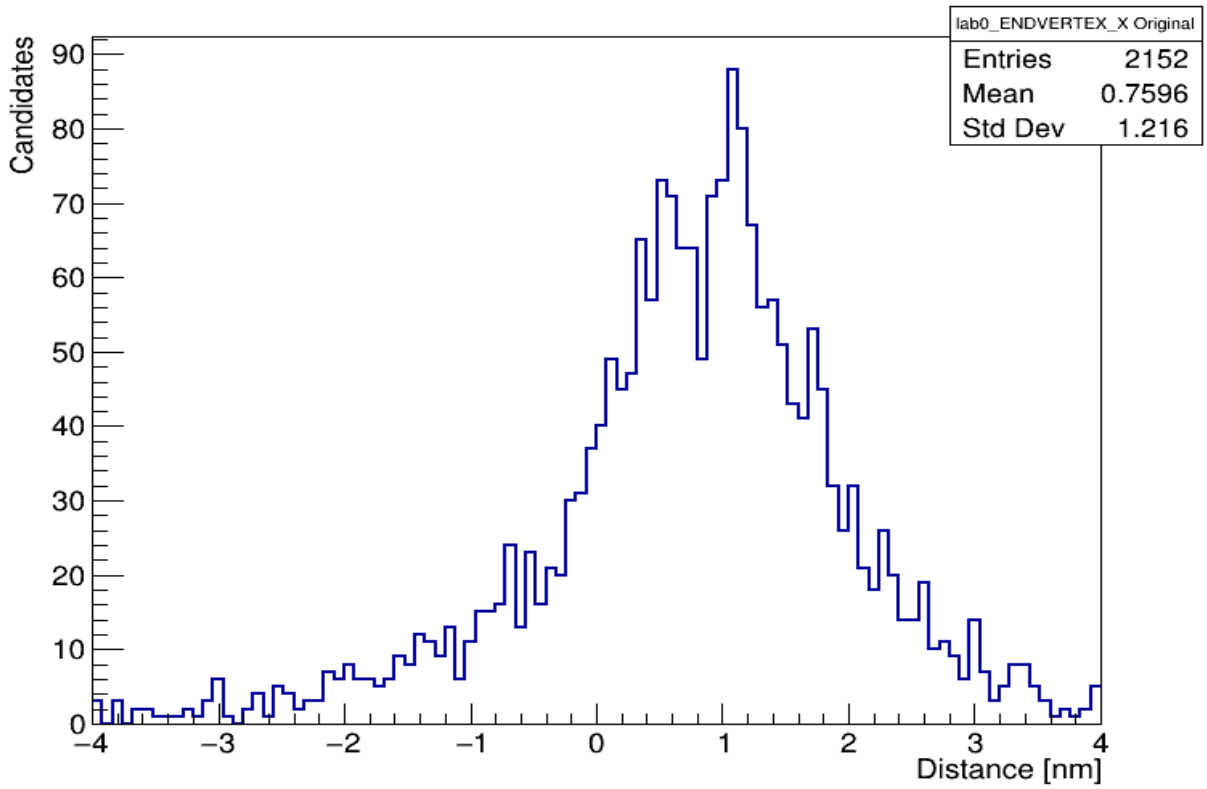


Rysunek 37 - Rozkład wartości residuum obliczonych dla współrzędnej Y wierzchołka rozpadu cząstki K_S^0 wraz z dopasowanym rozkładem normalnym

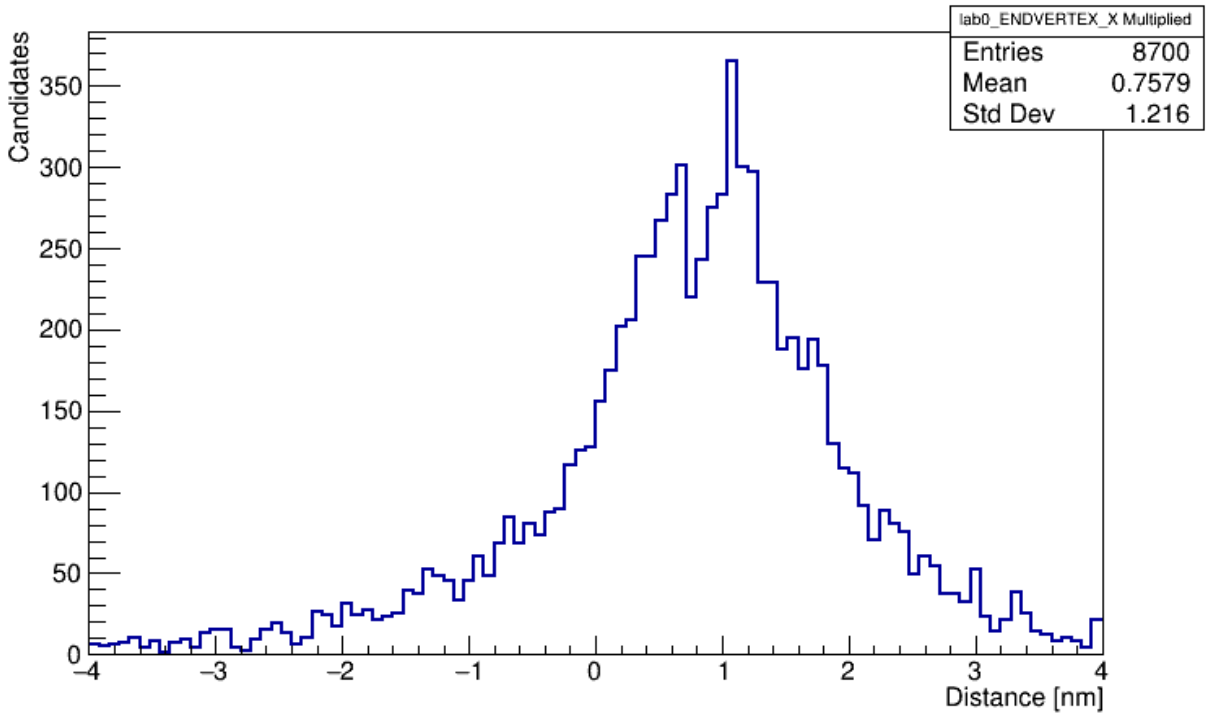


Rysunek 38 - Rozkład wartości residuum obliczonych dla współrzędnej Z wierzchołka rozpadu mezonu Ds wraz z dopasowanym rozkładem normalnym

Wszystkie przedstawione powyżej histogramy zawierają wartości oscylujące wokół zera. Odchylenie standardowe stanowi mały procent oryginalnych wartości użytych do obliczenia residuum (przedstawionych poniżej na rysunkach 39, 41, 43), co jest zgodne z założeniem dotyczącym rozkładu residuum przedstawionym w rozdziale 5.2. Wyrysowane krzywe dopasowania rozkładu ujawniają rozkład normalny. Przedstawione histogramy uznane zostały za poprawną reprezentację efektów aparaturowych, więc - podobnie jak wcześniej - stworzono na ich podstawie generator modulacji efektów aparaturowych dla współrzędnych wierzchołków powstania i rozpadu cząstek. Rysunki 40, 42, 44 przedstawiają powielone rozkłady omawianych przykładowych współrzędnych, uzyskane w wyniku potraktowania rozkładu oryginalnego generatorem modulacji efektów aparaturowych.

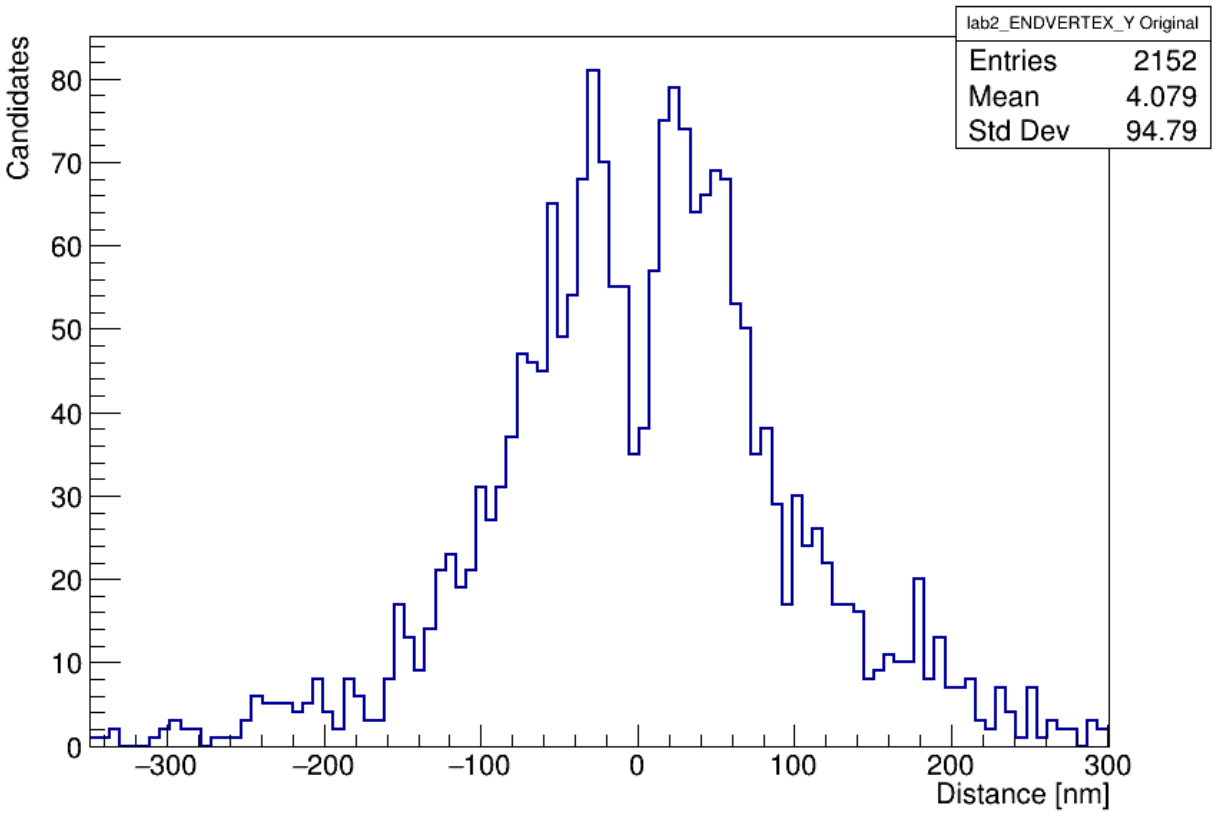


Rysunek 39 - Rozkład wartości współrzędnej X wierzchołka rozpadu mezonu Bs z próbki pierwotnej

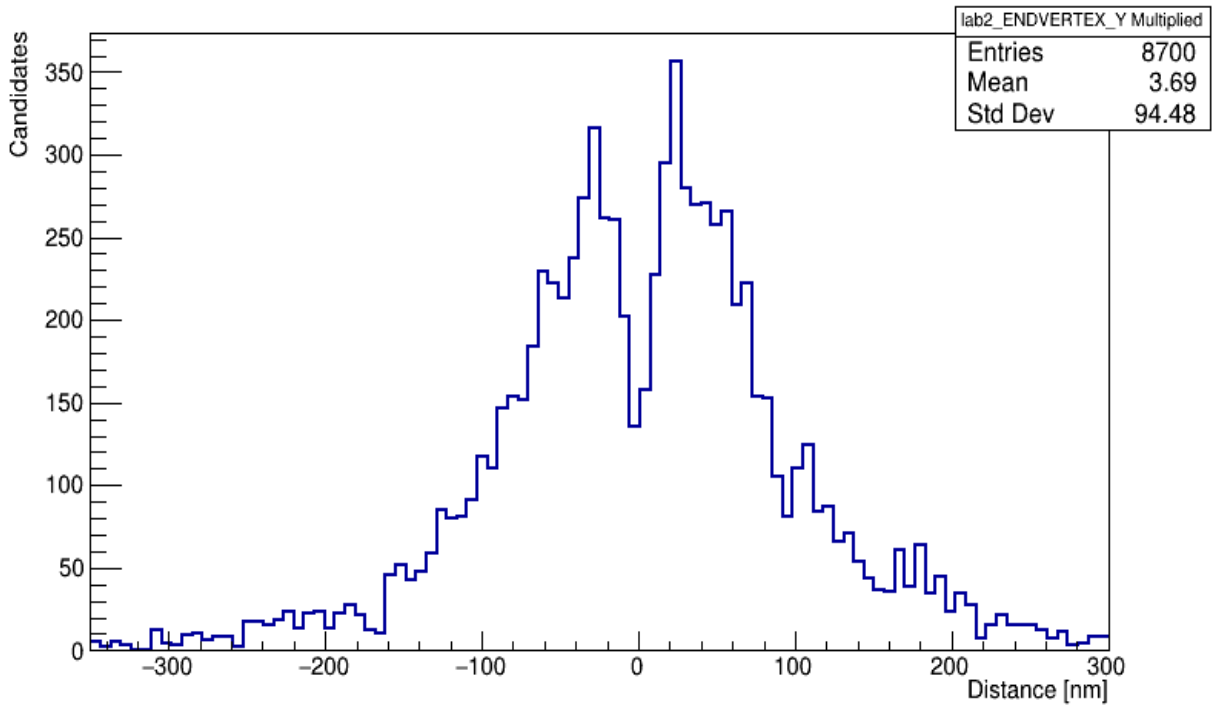


Rysunek 40 - Rozkład wartości współrzędnej X wierzchołka rozpadu mezonu Bs z próbki powielonej

W przypadku mezonu B_S wartości współrzędnej X wierzchołka rozpadu z próbki pierwotnej oraz zmultiplikowanej są niemalże identyczne. Różnica w wartości średniej widoczna jest na 4 miejscu znaczącym, zaś odchylenie standardowe jest identyczne. Rzeczą najwyraźniej świadczącą o różnicach między oboma próbkami są histogramy - od razu widać, że choć ich ogólny kształt jest taki sam, to występują różnice w wartościach "pików". Jest to dokładnie efekt, na którego uzyskanie liczono implementując algorytm multiplikujący. W przypadku pozostałych powielanych cech sytuacja wygląda podobnie. Rysunki 41, 42 przedstawiają odpowiadające histogramy dla współrzędnej Y wierzchołka rozpadu cząstki K_S^0.

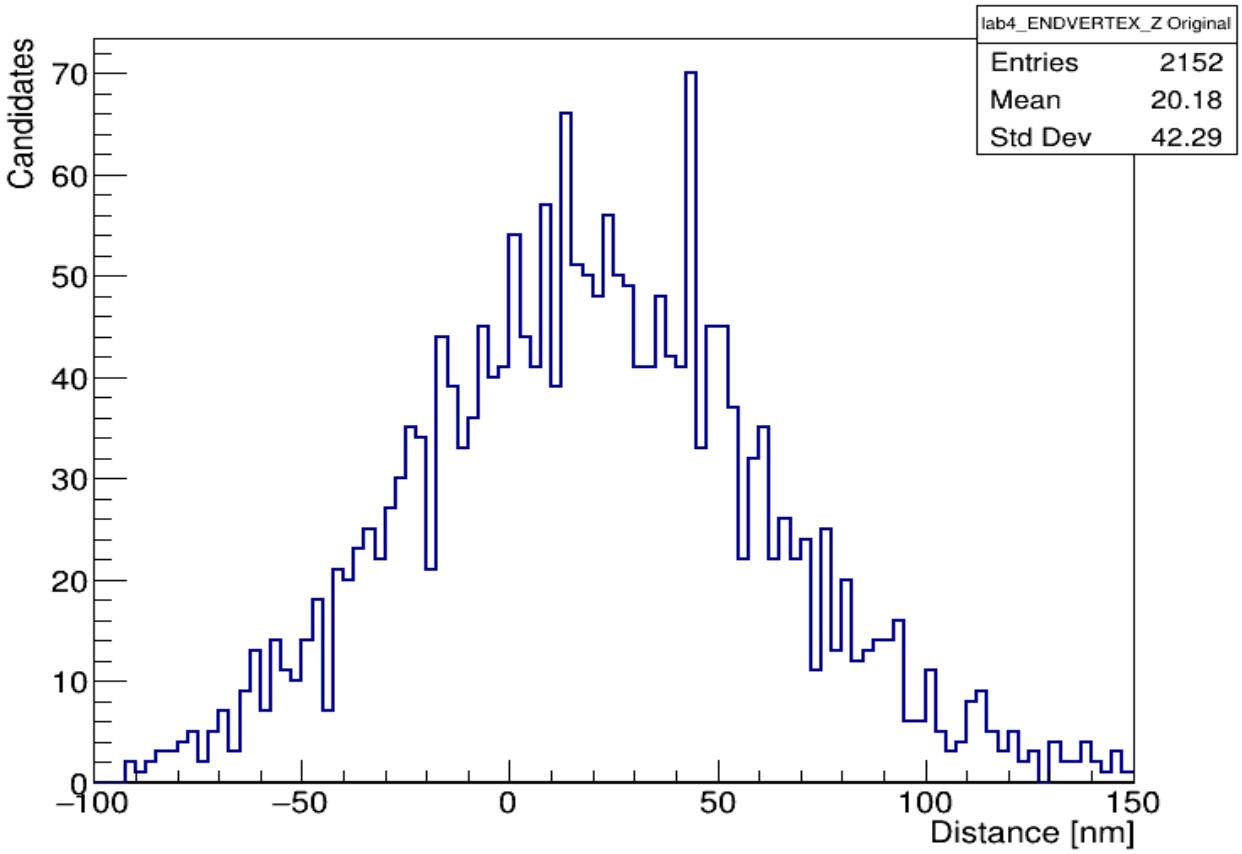


Rysunek 41 - Rozkład wartości współrzędnej Y wierzchołka rozpadu cząstki K_S^0 z próbki pierwotnej

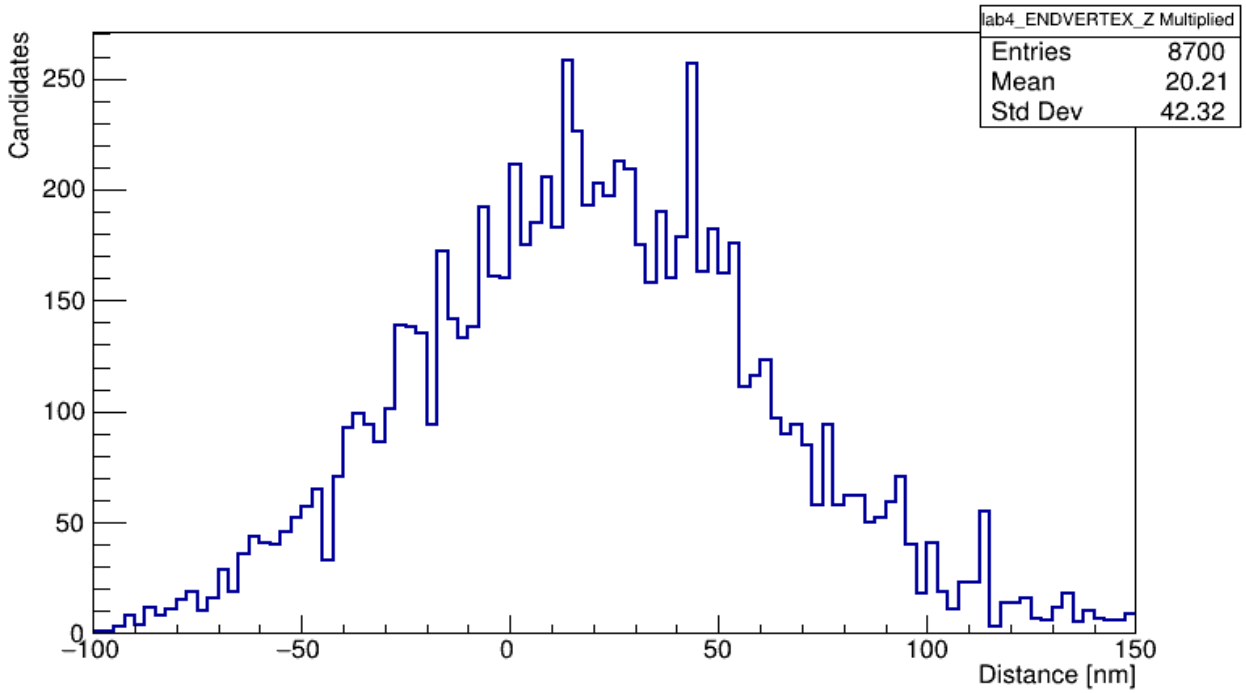


Rysunek 42 - Rozkład wartości współrzędnej Y wierzchołka rozpadu cząstki K_S^0 z próbki powielonej

Choć wartość średnia i odchylenie standardowe nie są w tym przypadku identyczne, są one bardzo zbliżone, a w skali wartości rejestrowanych na obydwu histogramach można uznać je za takie same. Tak jak w poprzednim przypadku, rozkład wartości na obu histogramach różni się jedynie w szczegółach pików, więc i tu można stwierdzić, że powielenie przebiegło zgodnie z oczekiwaniami. Ostatecznego potwierdzenia prawidłowości zaimplementowanej metody dostarcza trzecia omawiana w tej części cecha - współrzędna Z wierzchołka rozpadu mezonu D_S, której rozkłady przedstawiono na rysunkach 43, 44.



Rysunek 43 - Rozkład wartości współrzędnej Z wierzchołka rozpadu cząstki Ds z próbki pierwotnej

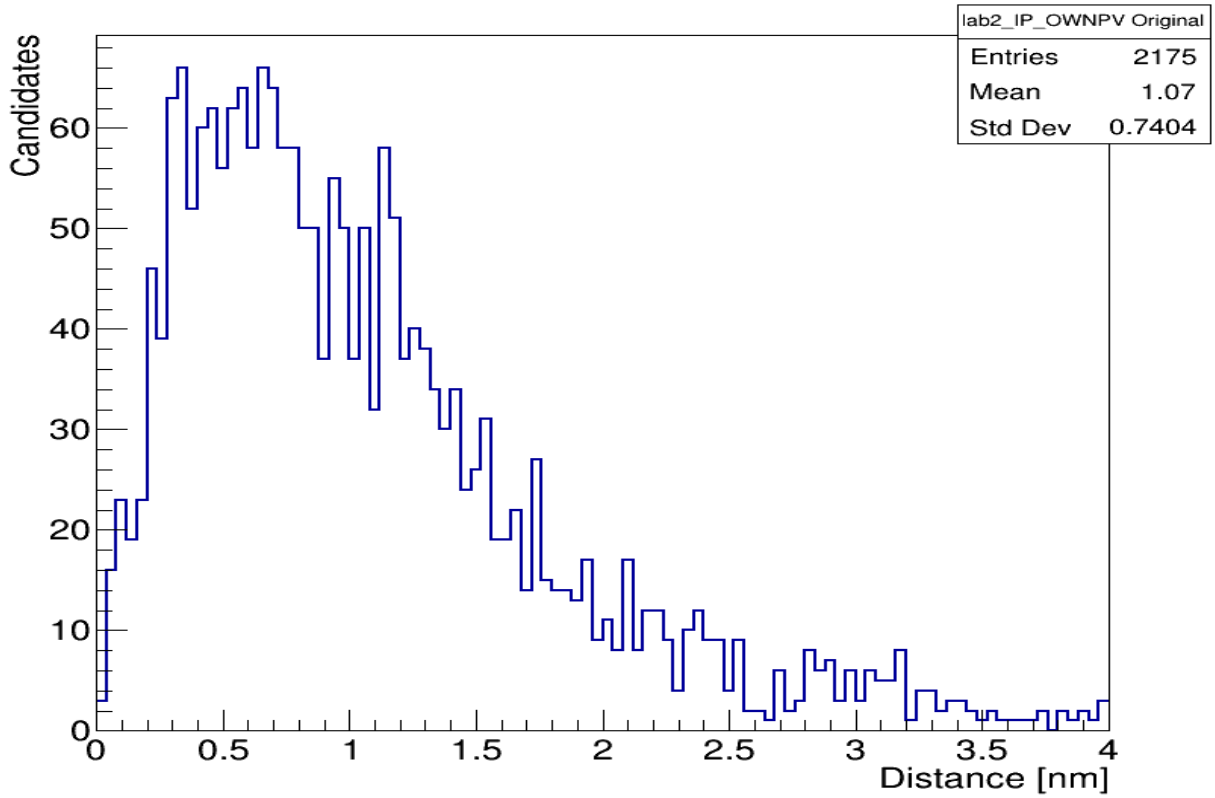


Rysunek 44 - - Rozkład wartości współrzędnej Z wierzchołka rozpadu cząstki Ds z próbki powielonej

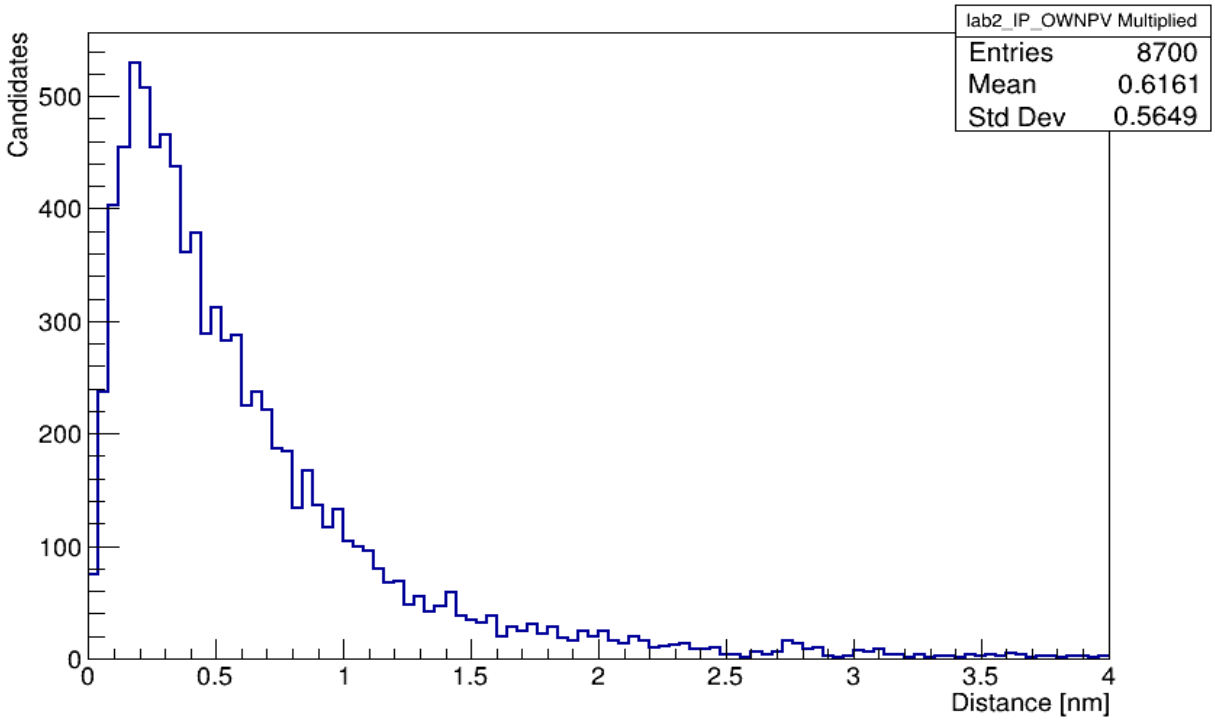
Podobnie jak w poprzednich dwóch przypadkach, powielony zestaw wartości cechy różni się od rozkładu oryginalnego niewielkimi szczegółami. Na podstawie tych analiz uznano, że działanie algorytmu multiplikującego stosowanego dla współrzędnych powstania i rozpadu wierzchołków cząstek jest prawidłowe, a uzyskane w wyniku jego działania wartości nadają się do wykorzystania w klasyfikatorze.

6.4. Geometryczny Parametr zderzenia

Na wstępie należy przypomnieć, że działanie algorytmu multiplikującego w części dotyczącej parametru zderzenia rózni się znacząco od wcześniejszych przypadków. Algorytm zamiast operować na wartościach parametru zderzenia "przed symulowaną detekcją" i po niej, bazuje na wartościach otrzymanych w wyniku powielenia wierzchołków powstania i rozpadu cząstek. Z tego względu spodziewano się, że odwzorowanie rozkładów pomiędzy próbką pierwotną a powieloną może nie być idealne. Należało jednak ocenić, czy otrzymane wartości można uznać za poprawne z fizycznego punktu widzenia, i czy użycie ich w klasyfikatorze BDT jest uzasadnione. Uzyskane rozkłady powielone wraz z odpowiadającymi im rozkładami oryginalnymi przedstawione zostały na rysunkach 45 - 48.



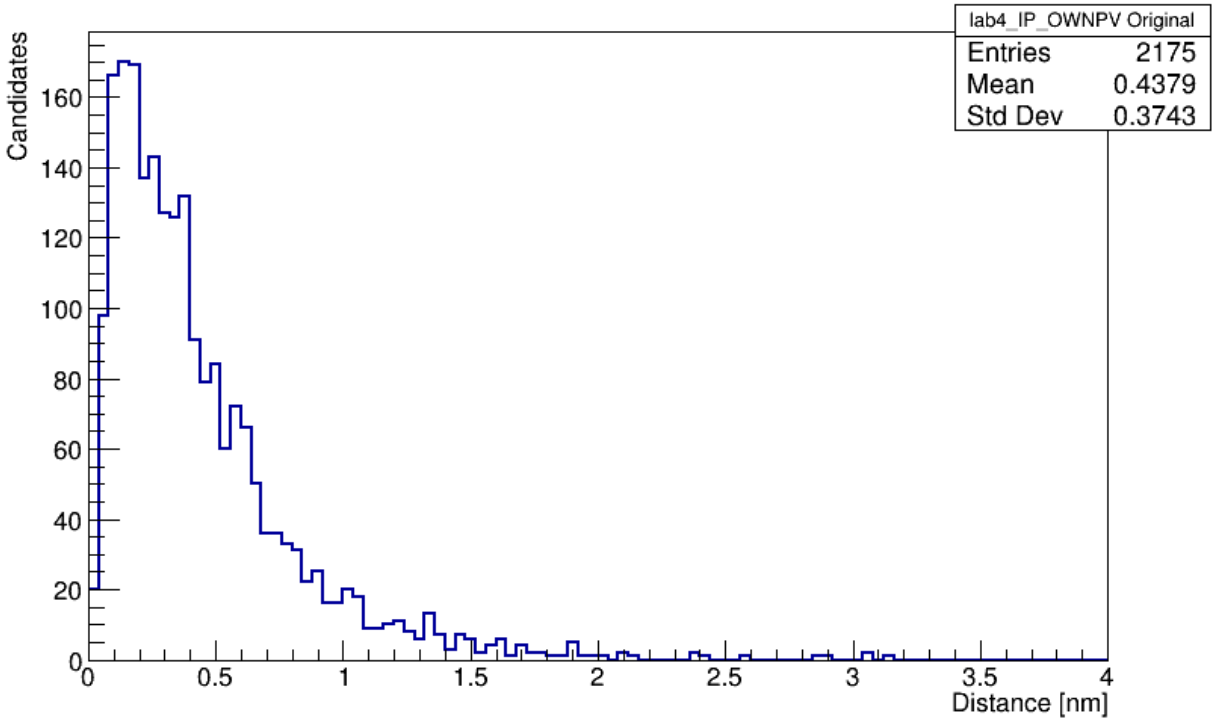
Rysunek 45 - Rozkład wartości parametru zderzenia dla cząstki K_S^0 w próbce pierwotnej



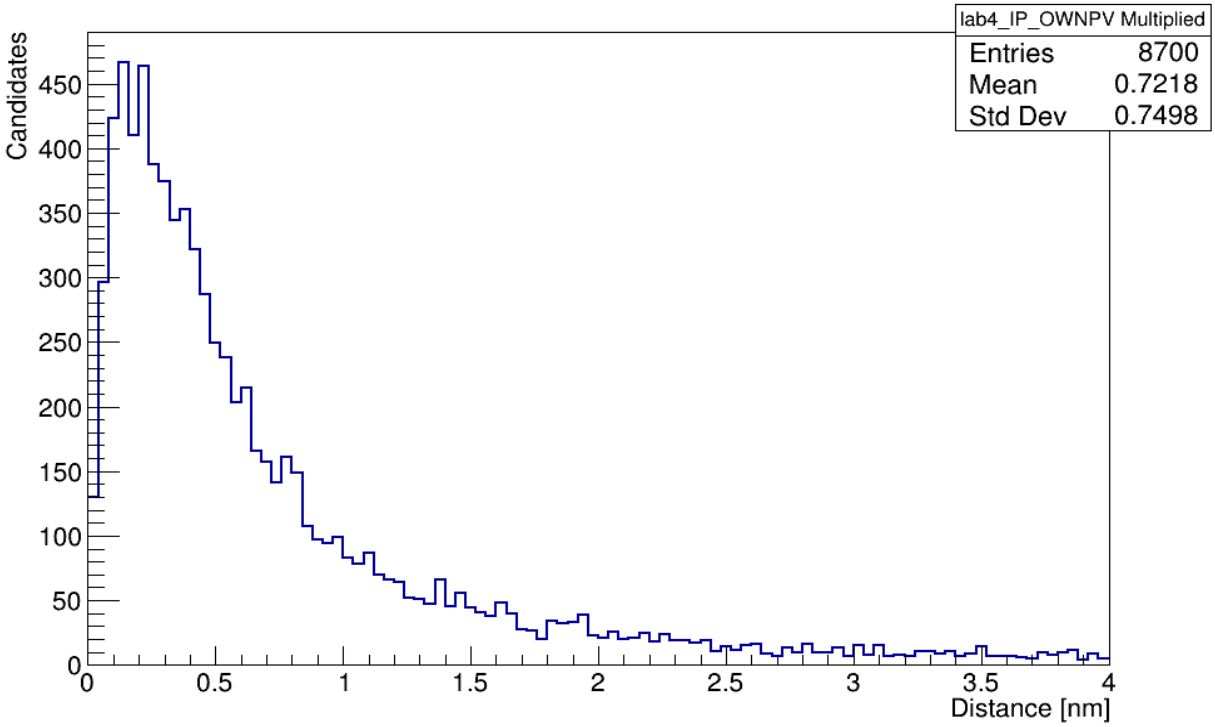
Rysunek 46 - Rozkład wartości parametru zderzenia dla cząstki K_S^0 w próbce powielonej

W przypadku cząstki K_S^0 istnienie prognozowanej różnicy między rozkładem pierwotnym i powielonym jest zauważalne na podstawie parametrów statystycznych - zarówno wartość średnia jak i odchylenie standardowe próbki multiplikowanej jest mniejsze, co skutkuje widocznym spłaszczeniem histogramu wzdłuż osi X. Histogram wartości oryginalnych zawiera stosunkowo więcej przypadków w przedziale wartości (0,0 - 1,0), następnie zaczyna opadać hiperbolicznie, rejestrując pojedyncze przypadki w odległości ok 3,5 nm i dalej. Z kolei w histogramie wartości powielonych widać wyraźny pik na odległości ok. 0,25 nm, gdzie rozpoczyna się hiperbola, osiągająca pojedyncze przypadki w odległości ok. 3 nm. Co ciekawe, na obu histogramach widoczny jest skokowy wzrost rejestrowanych wartości pomiędzy 2,7 a 3,2 nm.

W przypadku mezonu $\boldsymbol{D_S}$ sytuacja wygląda inaczej. Rozkłady parametru zderzenia dla tej cząstki zostały przedstawione na rysunkach 47 oraz 48.



Rysunek 47 - Rozkład wartości parametru zderzenia dla mezonu Ds w próbce pierwotnej



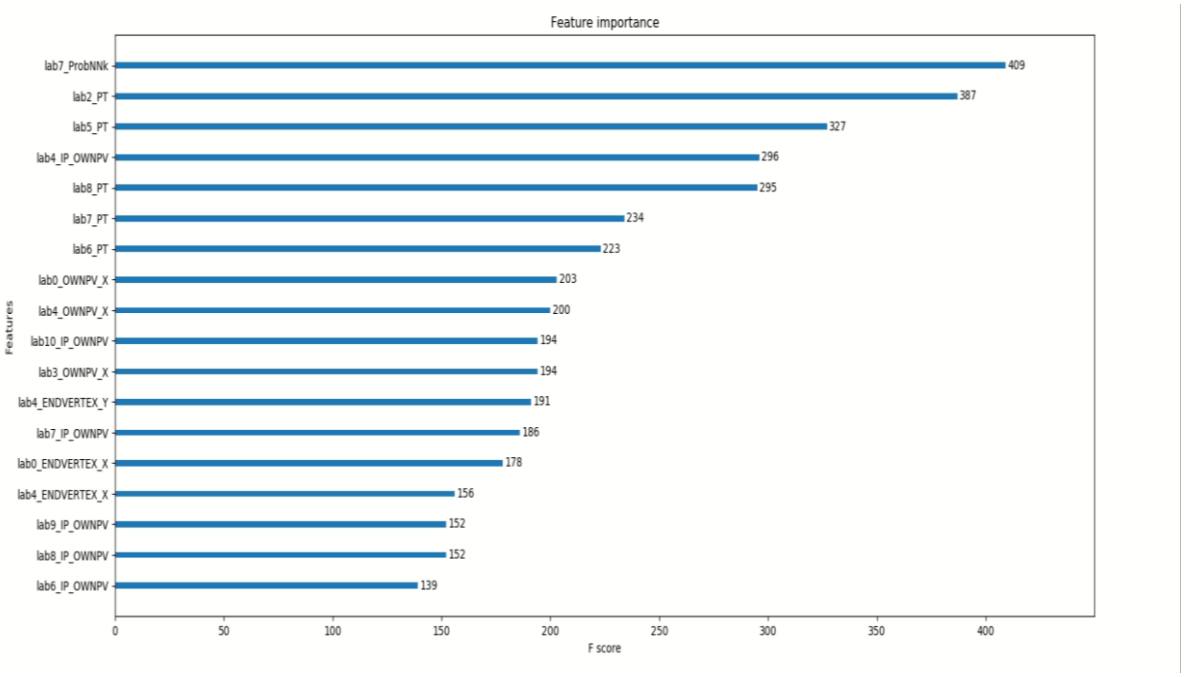
Rysunek 48 - Rozkład wartości parametru zderzenia dla mezonu Ds w próbce powielonej

Obserwując histogramy wyrysowane dla mezonu D_S łatwo zauważyć, że tym razem doszło do rozciągnięcia rozkładu powielonego względem osi X. Zarówno odchylenie standardowe jak i wartość średnia są tu większe, niemniej w kształcie histogramów różnice nie są wyraźne. Można jedynie stwierdzić, że wartości rozkładu pierwotnego, po osiągnięciu maksimum dla odległości ~0,3 nm szybko opadają w dół, aż do wartości 0,75 nm, a następnie da się zauważyć stopniowy spadek wartości, osiągający 0 w odległości 2,2 nm. Powyżej tej wartości, na histogramie rejestrowane są jedynie okazyjne, pojedyncze przypadki. Z kolei rozkład powielony, po osiągnięciu maksimum w podobnej odległości co rozkład oryginalny, zaczyna opadać bardziej hiperbolicznie, aż do osiągnięcia odległości około 4 nm. Da się zauważyć niewielką liczbę przypadków rejestrowanych w odległościach powyżej 3,2 nm, w których dla próbki pierwotnej żadne przypadki nie występowały.

6.5. Testowanie powielonych przypadków klasyfikatorem BDT

Klasyfikator trenowany próbką bazową

W przypadku tego klasyfikatora wejściowy zbiór danych liczący 4086 rekordów został podzielony na 3064 przypadki treningowe, oraz 1022 przypadki testowe. Liczba przypadków sygnałowych w zbiorze uczącym wynosiła 587. Program ustalił ważność (*Feature Importance*) cech cząstek wytypowanych do klasyfikatora w sposób przedstawiony na rysunku 49.

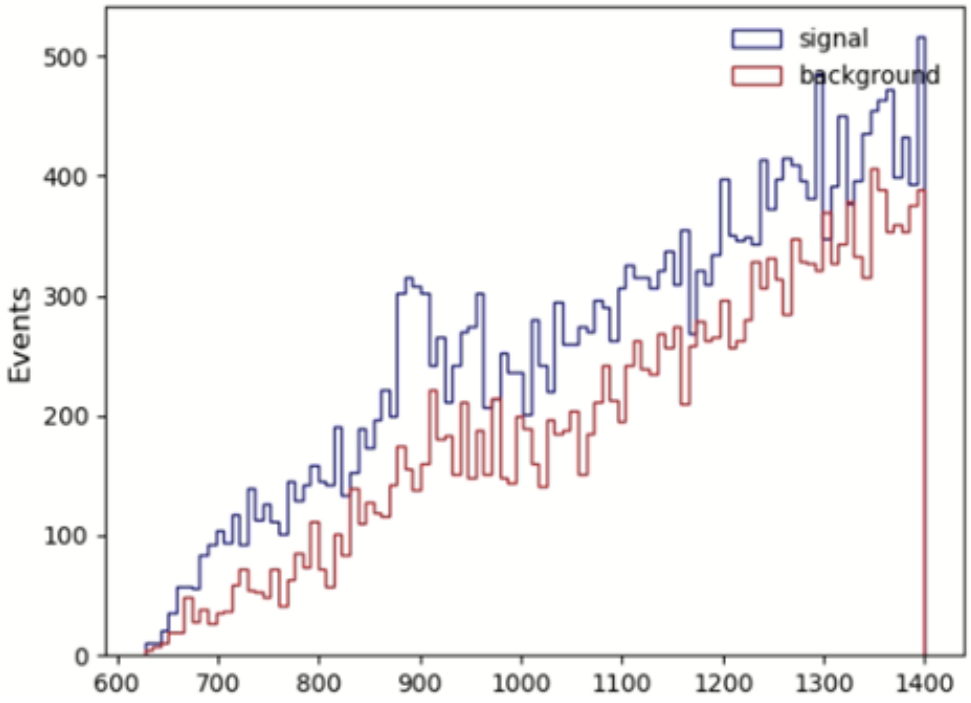


Rysunek 49 - Feature Importance cech cząstek wybranych do bazowego klasyfikatora BDT

Spośród zmiennych użytych w klasyfikatorze, pędy poprzeczne (PT) stanowią dla programu bardzo wartościową informację - aż 5 z 7 najważniejszych atrybutów stanowią pędy poprzeczne. Za najważniejszy atrybut program uznał *lab7ProbNNk* - jest to zmienna indentyfikacyjna wyznaczana na podstawie sieci neuronowej. W pierwszej dziesiątce znajdują się również dwie cechy związane z geometrycznym parametrem zderzenia (IP). Co ciekawe, 3 najmniej ważne cechy również stanowią dane dotyczące parametru zderzenia - ważność tej cechy jest więc silnie zależna od rodzaju cząstki. Współrzędne wierzchołków cząstek plasują się w większości w środku stawki *Feature Importance*. Warto podkreślić, że podział na atrybuty ważniejsze i

mniej ważne nie oznacza, że te drugie nie powinny być użyte w klasyfikatorze - wciąż dają one wartościowy wkład, jednak jest on mniejszy niż w przypadku cech najważniejszych.

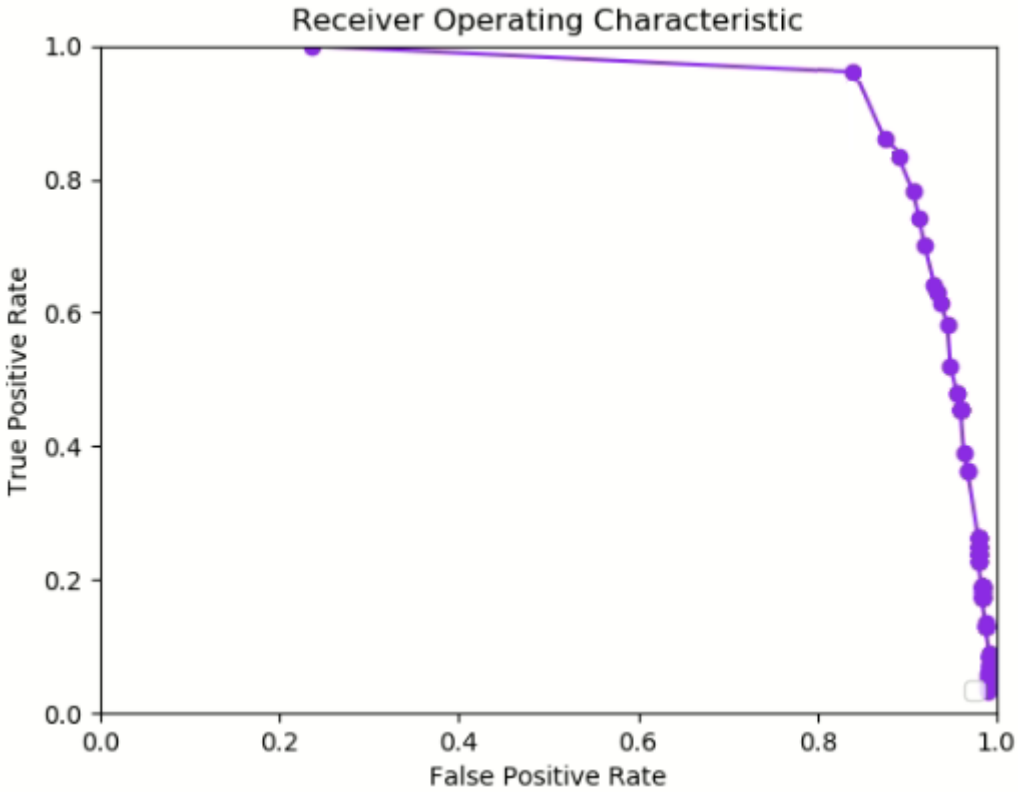
Do celów prezentacji efektywności klasyfikatora posłużono się przede wszystkim rozkładami wygenerowanymi dla masy cząstki K^* z uwagi na fakt, że kontrybucja sygnału i tła jest w tym przypadku wyraźnie widoczna i pozwala na określenie, czy zastosowana metoda powielania działa zgodnie z oczekiwaniami. Rozkład przypadków sygnałowych oraz tła dla wybranego przez wytrenowany program najlepszego cięcia BDT dla cząstki K^* przedstawia rysunek 50.



Rysunek 50 - Rozkład przypadków sygnałowych oraz tła dla optymalnego cięcia BDT dla masy cząstki K^* - klasyfikator wytrenowany zbiorem pierwotnym

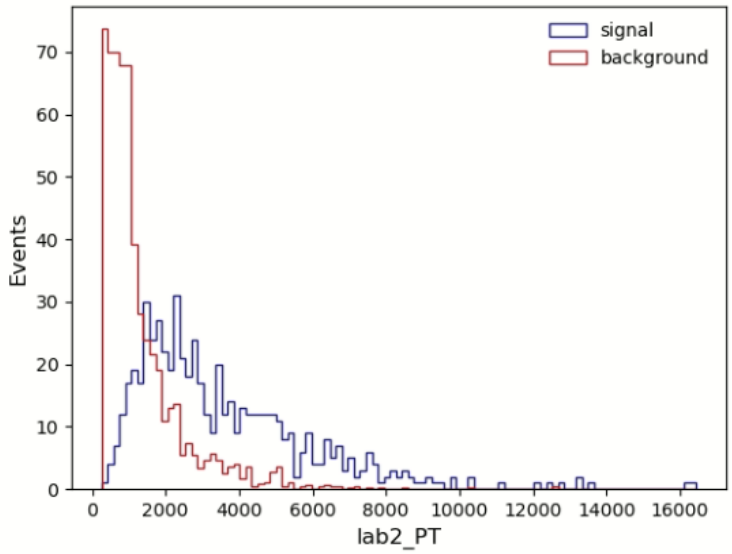
Powyższy rysunek stanowi pierwsze potwierdzenie, że pierwotny klasyfikator działa poprawnie - widoczny jest pik sygnałowy dla wartości 892 MeV, odpowiadającej wartości tablicowej. Pik ten nie jest widoczny dla przypadków tła. Powierzchnia pod krzywą ROC wykreśloną dla klasyfikatora wynosi 0,8428. Jest to wartość wysoka, co stanowi kolejne potwierdzenie dobrej jakości programu, jednak

wciąż pozostawiająca pole do poprawy. Krzywa ROC została przedstawiona na rysunku 51.

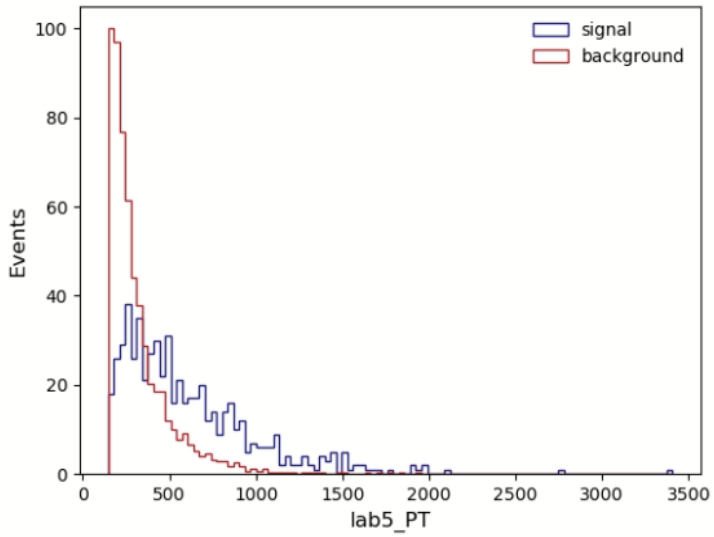


Rysunek 51 - Krzywa ROC dla klasyfikatora trenowanego niepowielonym zbiorem danych

Rysunki 52 - 55 przedstawiają przykładowe rozkłady zmiennych użytych w procesie trenowania klasyfikatora dla przypadków sygnału i tła. Rozkłady te stanowią uzasadnienie kolejności cech na wykresie *Feature Importance* - dla zmiennych o wysokiej ważności rozdział pomiędzy przypadkami sygnałowymi a tłem jest bardzo wyraźnie widoczny, zaś w przypadku cech z końca stawki, obydwa histogramy niemal się pokrywają.



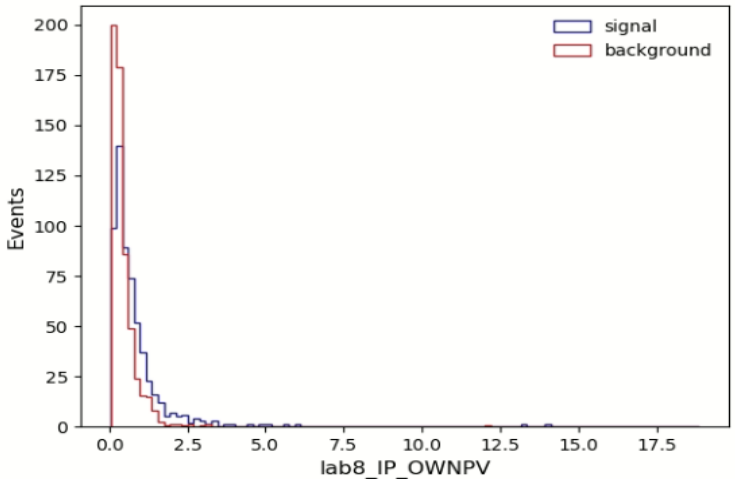
Rysunek 52 - Rozkład liczby przypadków sygnałowych oraz tła dla pędu poprzecznego cząstki K_S^0 wygenerowany przez bazowy klasyfikator



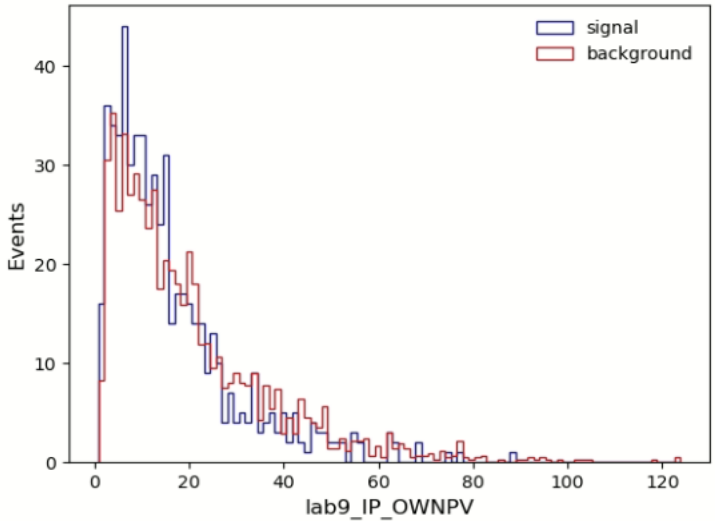
Rysunek 53 - Rozkład liczby przypadków sygnałowych oraz tła dla pędu poprzecznego cząstki γ wygenerowany przez bazowy klasyfikator

Obydwie prezentowane powyżej cechy są cechami ze szczytu listy *Feature Importance.* W obu przypadkach można zaobserwować wyraźnie zaznaczony pik liczby przypadków tła w początkowym przedziale obserwowanych wartości. Pik ten nie występuje w rozkładzie sygnałowym. Dla większych wartości pędów poprzecznych tło rejestrowane jest w kilkukrotnie mniejszej liczbie niż przypadki

sygnałowe. Złożenie tych dwóch zjawisk daje w efekcie histogramy wyraźnie od siebie oddzielone.



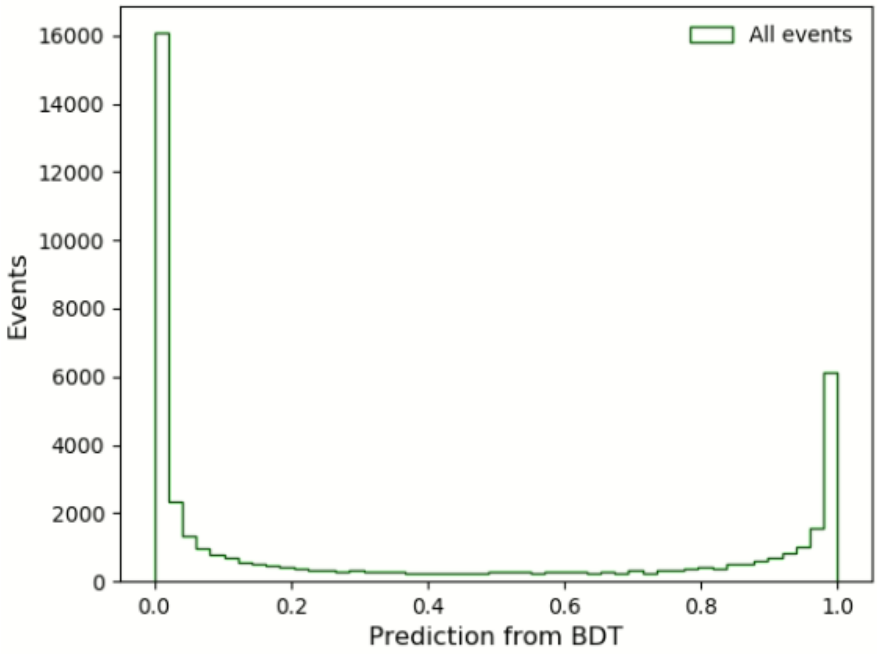
Rysunek 54 - Rozkład liczby przypadków sygnałowych oraz tła dla parametru zderzenia cząstki $K^{\pm}$ / $\pi^{\pm}$ wygenerowany przez bazowy klasyfikator



Rysunek 55 - Rozkład liczby przypadków sygnałowych oraz tła dla parametru zderzenia cząstki $\pi^{\pm}$ pochodzącej z rozpadu K_S^0 wygenerowany przez bazowy klasyfikator

W przypadku dwóch powyższych rozkładów różnice między histogramami są nieznaczne. Obserwowane piki występują zarówno dla histogramów przypadków sygnałowych, jak i dla tła, zakresy osiąganych wartości są bardzo zbliżone, a przebiegi niemal się pokrywają. Z tego powodu cechy te zostały ocenione jako mało znaczące w procesie klasyfikacji.

Po zakończeniu fazy nauki, do programu dostarczono zupełnie nowy zbiór danych rzeczywistych zebranych w detektorze LHCb, celem jego klasyfikacji. Odpowiedź BDT dla tego zbioru przedstawia rysunek 56.

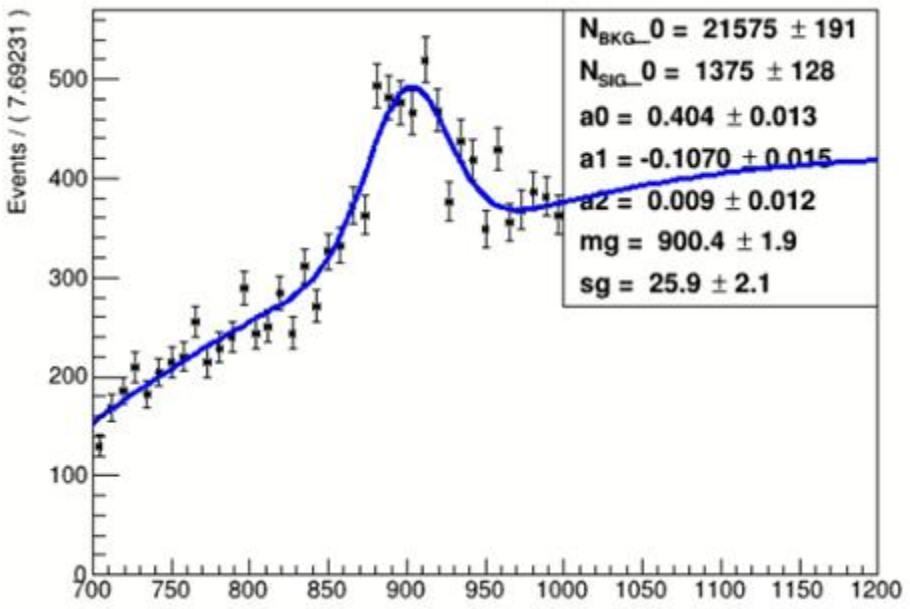


Rysunek 56 - Rozkład odpowiedzi BDT dla zbioru danych poddanego klasyfikacji.

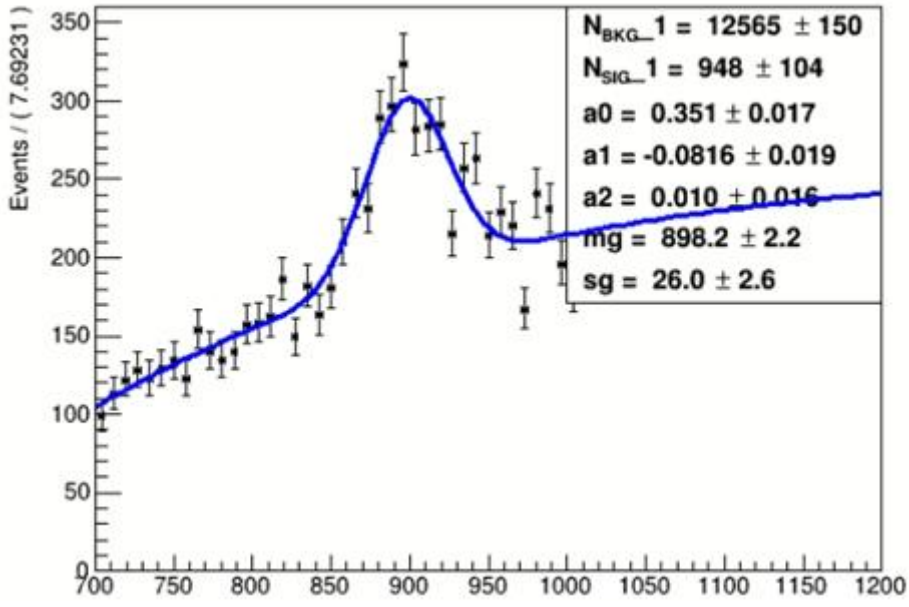
Na powyższym rozkładzie łatwo zauważyć dwa piki, zlokalizowane na dwóch krańcach przedziału wartości odpowiedzi BDT (0,0 - 0,2 oraz 0,8 - 1,0). Pomiędzy nimi, w przedziale 0,2 - 0,8 liczba zarejestrowanych zdarzeń jest niewielka i prawie stała.

Kolejnym krokiem niezbędnym do zakończenia klasyfikacji było dokonanie cięcia, czyli podzielenie histogramu odpowiedzi BDT na dwie części - jedną zawierającą wszystkie przypadki sygnałowe, oraz drugą zawierającą wszystkie przypadki tła. Rysunek 57 przedstawia rozkład przypadków sygnałowych masy cząstki K^* bez cięcia BDT, zaś na rysunkach 58 - 60 przedstawione zostały wygenerowane przez klasyfikator rozkłady przypadków sygnałowych tej zmiennej dla różnych wartości cięcia. Wszystkie rozkłady uzupełnione zostały o krzywe dopasowania rozkładu prawdopodobieństwa, zaznaczone niebieskimi liniami. Na rysunkach widać wyraźnie, że krzywe te są złożeniem dwóch rozkładów - rozkładu

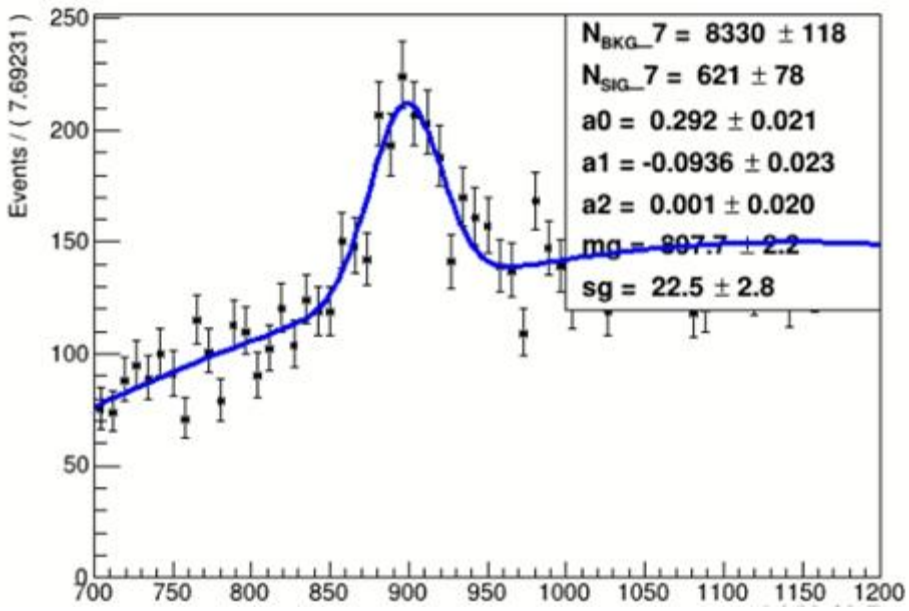
Gaussa modelującego przypadki sygnałowe, oraz krzywej wielomianowej modelującej tło. Klasyfikator wyliczył również liczby przypadków zawartych pod każdą z krzywych, co posłużyło do wyznaczenia wielkości *Figure of Merit*.



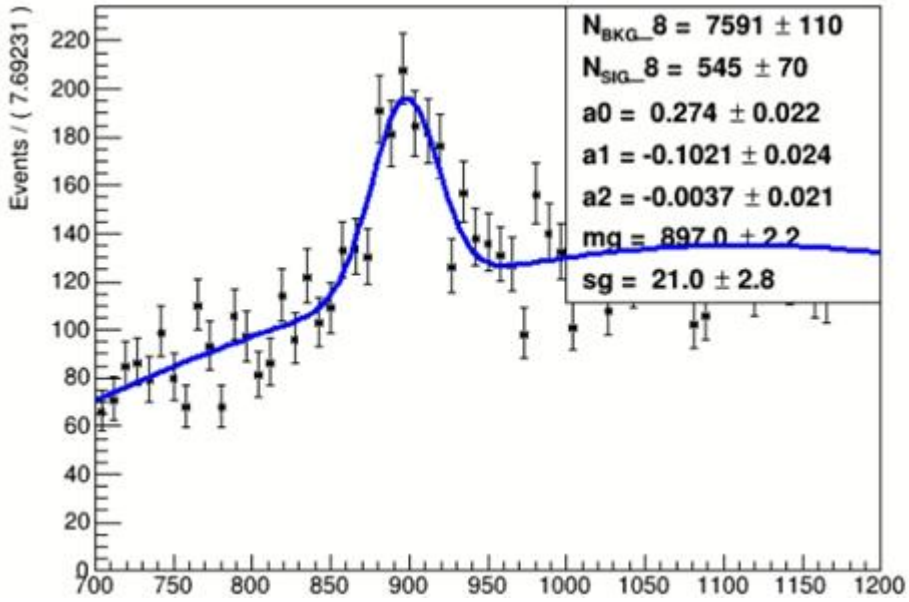
Rysunek 57 - Rozkład przypadków sygnałowych wartości masy cząstki K^* bez cięcia BDT



Rysunek 58 - Rozkład przypadków sygnałowych wartości masy cząstki K^* dla cięcia BDT = 0,2



Rysunek 59 - Rozkład przypadków sygnałowych wartości masy cząstki K^* dla cięcia BDT = 0,4



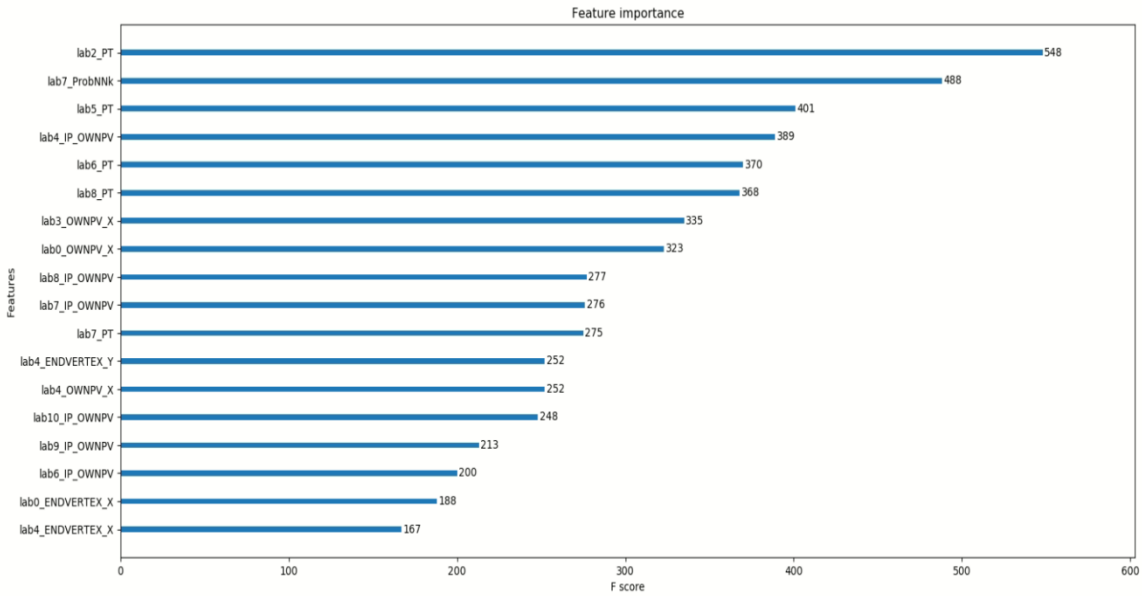
Rysunek 60 - Rozkład przypadków sygnałowych wartości masy cząstki K^* dla cięcia BDT = 0,6

Pomijając różnice w liczbie rejestrowanych zdarzeń, powyższe wykresy wyglądają identycznie do tego stopnia, że niemożilwe jest określenie, dla którego cięcia klasyfikacja daje najlepsze efekty. Pomocą w tej sytuacji jest omawiana wcześniej wielkość *Figure of Merit.* Dysponując przedstawioną na wykresach wiedzą

o liczbie przypadków zaklasyfikowanych jako sygnał i jako tło, możliwe stało się wyliczenie *Figure of Merit*. W tej chwili wielkości te nie niosą ze sobą wartościowej informacji - ich znaczenie uwidoczni się w momencie porównania ich z wartościami FoM dla klasyfikatora trenowanego próbką powieloną. Porównanie FoM wyznaczonych dla każdego z cięć dla obu klasyfikatorów przedstawia tabela 1

Klasyfikator trenowany próbką powieloną

W przypadku tego klasyfikatora wejściowy zbiór danych liczący 7787 rekordów został podzielony na 5988 przypadków treningowych, oraz 1799 przypadków testowych. Liczba przypadków sygnałowych w zbiorze uczącym wynosiła 2923. Ważność (*Feature Importance*) cech cząstek wytypowanych do klasyfikatora została ustalona w sposób przedstawiony na rysunku 61.

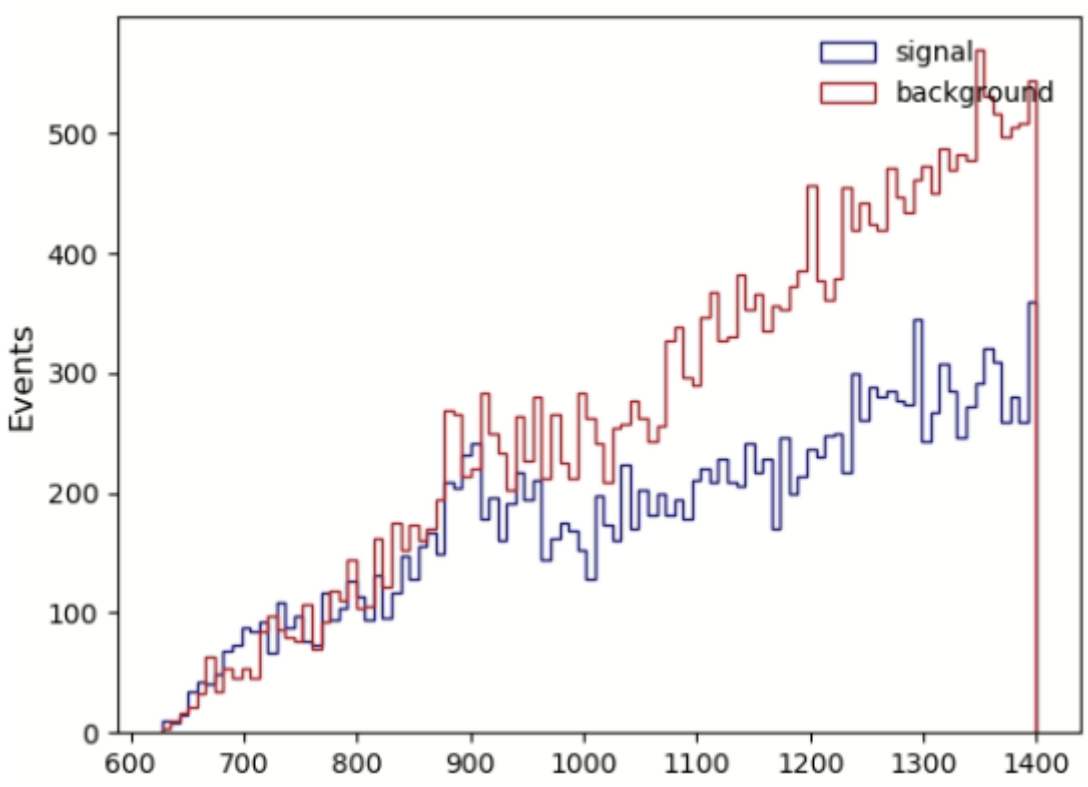


Rysunek 61 - *Feature Importance* cech cząstek wybranych do klasyfikatora BDT opartego o powielony zbiór danych

W stosunku do rozkładu *Feature Importance* dla bazowego klasyfikatora, można zauważyć wzrost znaczenia atrybutów związanych z pędem poprzecznym. Podobnie jak poprzednio zajmują one czołowe miejsca, jednak ich wartości *F score*

są wyraźnie wyższe. Do zmian doszło na najmniej ważnych pozycjach - końcowe miejsca przypadły tym razem cechom związanym ze współrzędnymi wierzchołków powstania i rozpadu cząstek. Ważność zmiennych reprezentujących parametr zderzenia, podobnie jak w poprzednim przypadku, zależy od rodzaju cząstki.

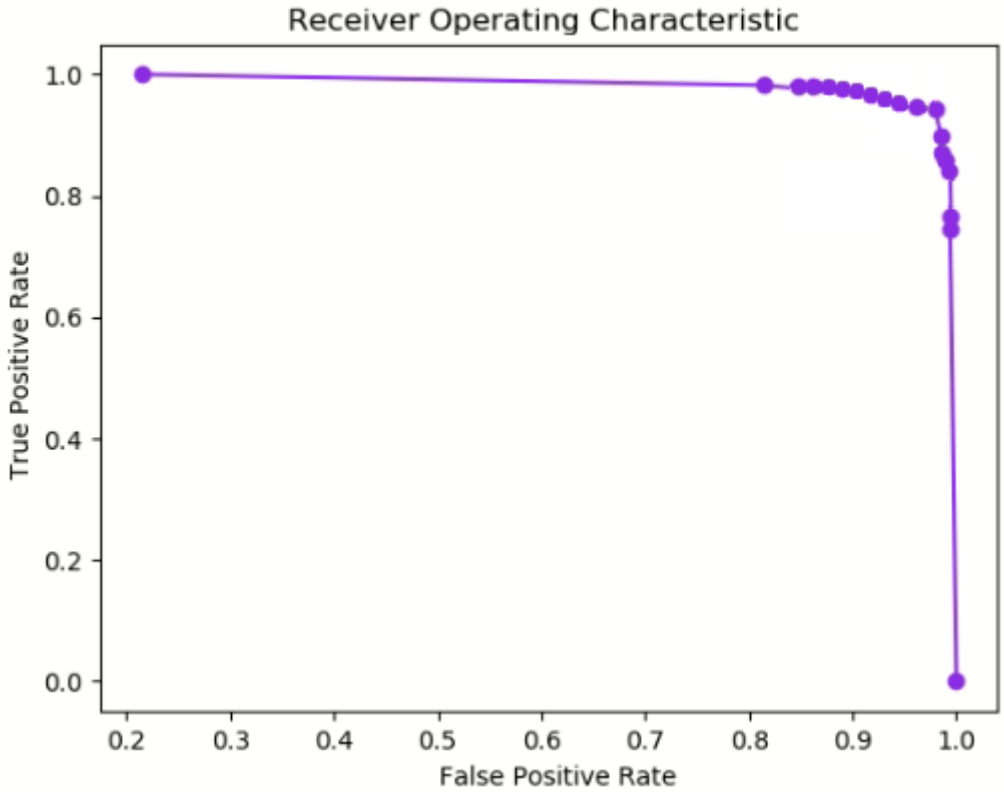
Tak jak poprzednio, prezentacja wyników klasyfikacji zostanie przedstawiona przede wszystkim w oparciu o rozkłady wygenerowane dla masy cząstki K^*. Rozkład przypadków sygnałowych oraz tła dla wybranego przez program najlepszego cięcia BDT dla cząstki K^* przedstawia rysunek 62.



Rysunek 62 - Rozkład przypadków sygnałowych oraz tła dla optymalnego cięcia BDT dla masy cząstki K^* - klasyfikator uczony zbiorem powielonym

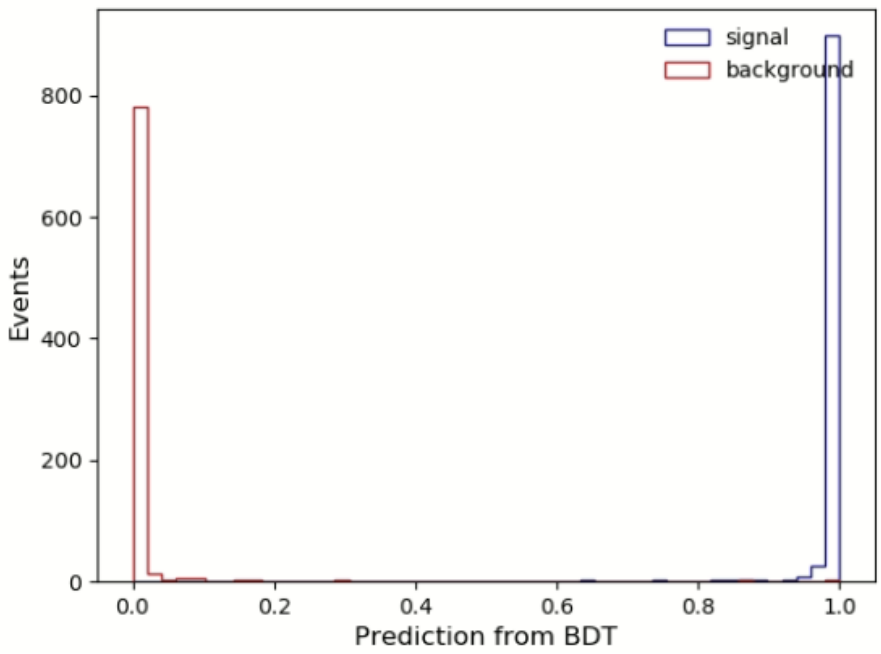
Dla klasyfikatora trenowanego przy pomocy danych powielonych, przy tak dobranym biningu, pik sygnałowy nie jest tak wyraźny jak w przypadku programu uczonego danymi bazowymi. Rzeczą zasługującą na uwagę jest natomiast bardzo dobra separacja przypadków sygnałowych od tła dla wysokich wartości masy. Powierzchnia pod krzywą ROC wykreśloną dla tego klasyfikatora wynosi 0,9322.

Wartość ta jest znacznie wyższa od tej otrzymanej dla poprzedniego programu, co jest pierwszą wskazówką, że omawiany obecnie klasyfikator działa lepiej. Krzywa ROC została przedstawiona na rysunku 63. W kwestii rozkładów zmiennych użytych w procesie trenowania klasyfikatora dla przypadków sygnału i tła sytuacja nie zmienia się względem poprzedniego programu - histogramy dla zmiennych o wysokim Feature Importance były wyraźnie rozdzielone, zaś te o małej ważności nakładały się na siebie.



Rysunek 63 - Krzywa ROC dla klasyfikatora trenowanego powielonym zbiorem danych

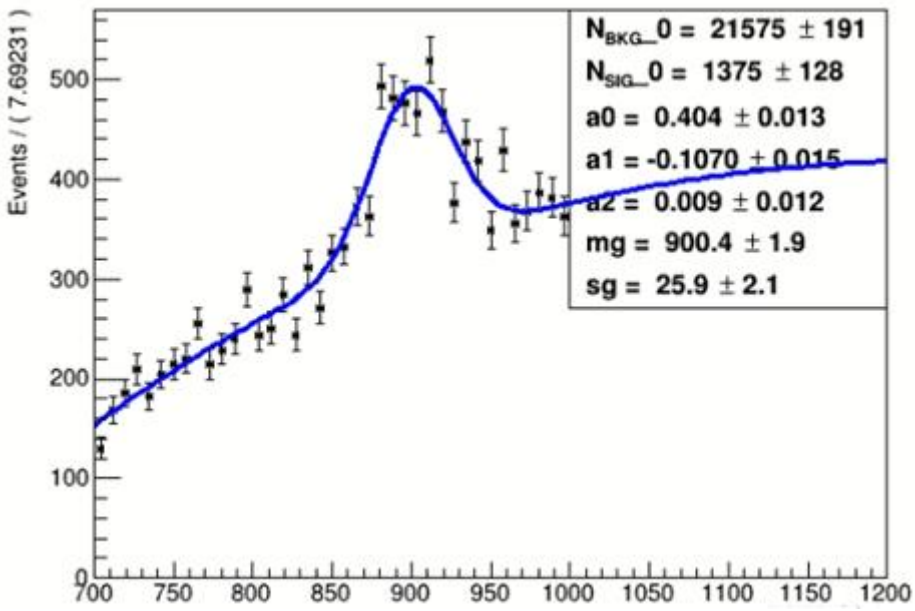
W kolejnym kroku - po zakończeniu fazy nauki - program otrzymał nowy, wcześniej nie widziany zbiór danych celem jego klasyfikacji. Rysunek 64 przedstawia rozkład odpowiedzi BDT dla przypadków rozpadu w testowym podzbiorze powielonego zbioru danych.



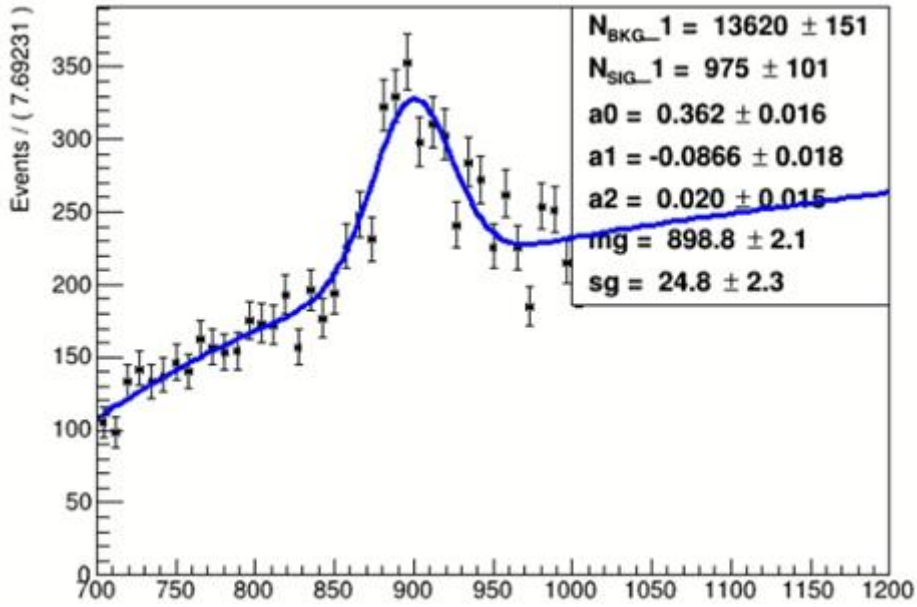
Rysunek 64 - Rozkład odpowiedzi BDT dla przypadków rozpadu w podzbiorze testowym powielonego zbioru danych

Porównanie powyższego histogramu z jego odpowiednikiem dla klasyfikatora bazowego pozwala zauważyć, że w programie trenowanym danymi powielonymi rozgraniczenie na dwa piki jest o wiele bardziej wyraźne. W poprzednim rozkładzie pewna ilość przypadków była rejestrowana w pełnym spektrum zakresu wartości odpowiedzi BDT. W obecnie analizowanym klasyfikatorze, niemal żadne rekordy nie są rejestrowane w przedziale 0,1 - 0,9. Obserwacja ta pozwala stwierdzić, że program jest w stanie z większą pewnością jednoznacznie przyporządkować przypadki.

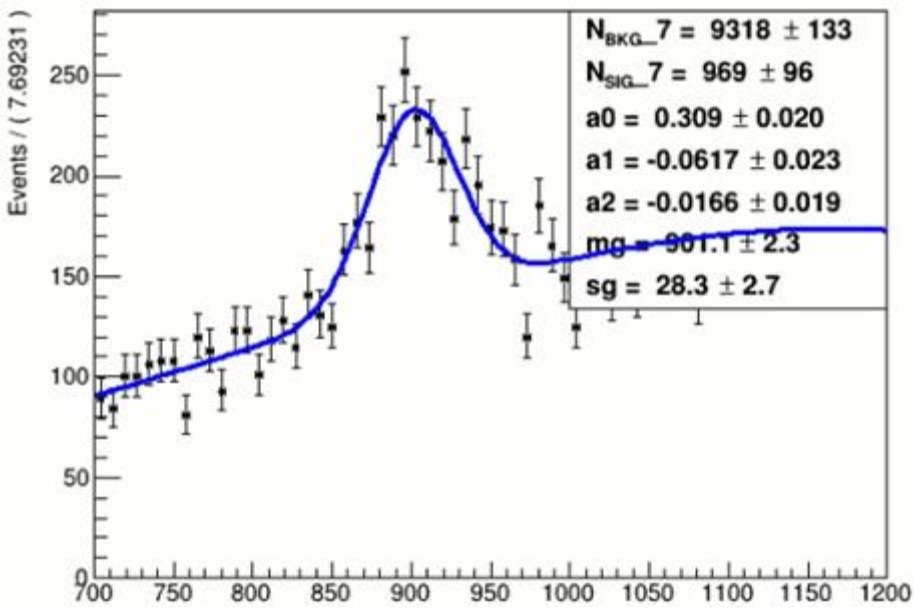
Znając histogram odpowiedzi BDT w całej klasyfikowanej próbce przystąpiono do analizy rozkładów przypadków sygnałowych masy cząstki K^* dla różnych cięć BDT. Rozkłady te przedstawiają rysunki 65 - 68. Tak jak w przypadku klasyfikatora bazowego, rysunki te zawierają krzywe dopasowania rozkładu prawdopodobieństwa, zaznaczone niebieskimi liniami.



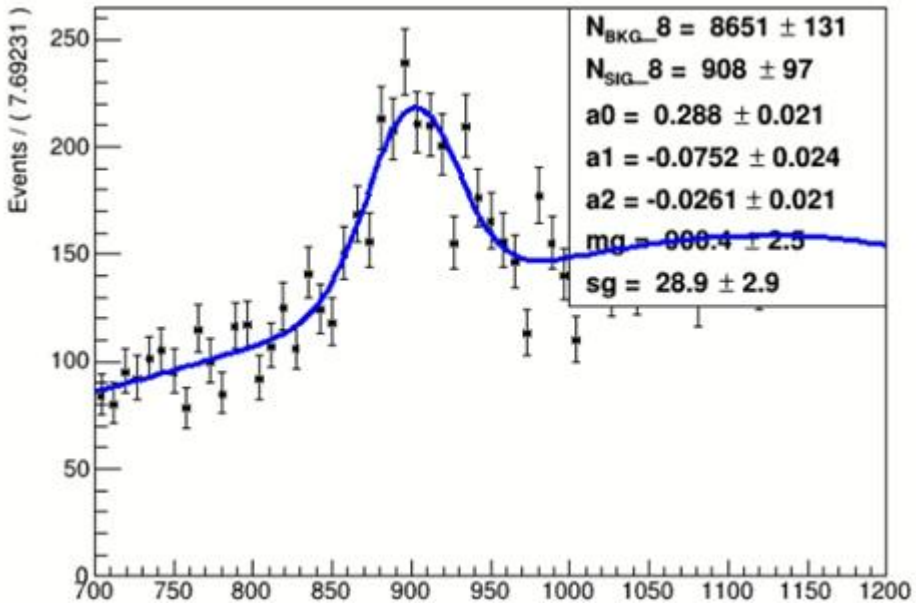
Rysunek 65 - Rozkład przypadków sygnałowych wartości masy cząstki K^* bez cięcia BDT



Rysunek 66 - Rozkład przypadków sygnałowych wartości masy cząstki K^* dla cięcia BDT = 0,2



Rysunek 67 - Rozkład przypadków sygnałowych wartości masy cząstki K^* dla cięcia BDT = 0,4



Rysunek 68 - Rozkład przypadków sygnałowych wartości masy cząstki K^* dla cięcia BDT = 0,6

Tak jak w przypadku pierwszego klasyfikatora, zwiększenie wartości cięcia powoduje każdorazowo zmniejszenie liczby rejestrowanych przypadków. Przebieg krzywych dopasowania rozkładu jest niemal identyczny w obu klasyfikatorach, więc

aby móc ocenić poprawę jakości klasyfikacji konieczne jest porównanie wielkości *Figure of Merit* dla obu programów.

Wartość cięcia BDT	Wartość FoM - klasyfikator bazowy	Wartosc FoM - klasyfikator powielony
0,0	9,0764	9,0764
0,2	8,1552	8,0705
0,4	6,5638	9,5539
0,6	6,0421	9,2871

Tabela 1 - Porównanie wartości Figure of Merit dla klasyfikacji masy cząstki K^* dla różnych cięć BDT dla obu klasyfikatorów

Porównanie wartości *Figure of Merit* obu programów wyraźnie pokazuje który klasyfikator jest dokładniejszy. Choć dla cięcia 0,2 można zauważyć pozorny spadek wydajności drugiego programu, to przy większych wartościach cięć wyższość programu uczonego powielonym zbiorem danych jest niezaprzeczalna - jego FoM jest wyższa o połowę dla cięć 0,4 oraz 0,6.

7. Wnioski

Analizując wyniki otrzymane w pierwszej części niniejszej pracy można stwierdzić, że autorski algorytm multiplikujący dane został zaimplementowany poprawnie. Wszystkie otrzymane histogramy powielanych wartości cech cząstek odwzorowują rozkłady z próbki bazowej nieznacznie je modyfikując - zgodnie z założeniem algorytmu, co pokazano na rysunkach 18 - 26, 30 - 35, 39 - 48, . Potencjalne korzyści płynące z tego faktu są oczywiste - możliwość generowania praktycznie dowolnej liczby poprawnych z fizycznego punktu widzenia rekordów przy znikomym nakładzie czasu i kosztów.

Jednak sam fakt powielenia danych w zgodny z fizycznymi prawidłami sposób nie miałby większego znaczenia z naukowego punktu widzenia, gdyby nie dało się

otrzymanych zbiorów wykorzystać w celu poprawy jakości analizy rzeczywistych danych przetwarzanych w eksperymencie LHCb. Celem drugiej części pracy było więc potwierdzenie przydatności wygenerowanych przypadków w realnym scenariuszu. Uzyskane efekty potwierdziły wartość naukową multiplikacji danych - wszystkie badane metryki są poprawne i wykazały znaczną poprawę jakości programu klasyfikującego uczonego przy pomocy powielonego zbioru danych. W przypadku krzywej ROC poprawa wyniosła około 11% (rysunki 51 oraz 63), zaś w przypadku *Figure of Merit* udało się otrzymać uzysk rzędu 40 - 50%, przedstawiony w tabeli 1.

8. Podsumowanie

Głównym celem niniejszej pracy było potwierdzenie przydatności autorskiej techniki powielania danych z eksperymentu LHCb dotyczących rzadkiego rozpadu $\boldsymbol{B_S^0 \to K^{*\pm} D_s^{*\pm}}$. W analizie skupiono się na zagadnieniu klasyfikacji obserwowanych przypadków na sygnał i tło kombinatoryczne. Na obecną chwilę można powiedzieć, że zastosowana technika daje bardzo obiecujące wyniki - we wszystkich fazach niniejszej pracy zakładano, że zaobserwowanie choć minimalnej poprawy będzie dużym sukcesem, jako że w zagadnieniu klasyfikacji nawet uzysk rzędu ułamka procenta jest znacznym osiągnięciem. Otrzymane wyniki przerosły oczekiwania - poprawa jakości rzędu kilkunastu procent jest bardzo dużym krokiem w stronę szerokiego zastosowania techniki multiplikacji danych w zagadnieniach z zakresu fizyki wysokich energii.

9. Źródła

[1] Andreas Kaplan; Michael Haenlein (2019) Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence, Business Horizons, 62(1), 15-25.

[2] https://ordergroup.co/en/blog/why-python-best-machine-learning (dostęp 3.11.2020)

[3] https://xgboost.readthedocs.io/en/latest/ (dostęp 3.11.2020)

[4] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ (dostęp 3.11.2020)

[5] https://iopscience.iop.org/article/10.1088/1748-0221/3/08/S08001 (dostęp 3.11.2020)

[6] https://edms.cern.ch/ui/file/973073/1/Report_on_080919_incident_at_LHC__2_.pdf (dostęp 3.11.2020)

[7] http://www.accelerators-for-society.org/about-accelerators/index.php?id=21 (dostęp 3.11.2020)

[8] J. Phys.: Conf. Ser. 878 012012

[9] Int. J. Mod. Phys. A 30, 1530022 (2015)

[10] https://lhcb.github.io/starterkit-lessons/first-analysis-steps/dataflow.html (dostęp 3.11.2020).

[11] https://www.jeremyjordan.me/decision-trees/ (dostęp 3.11.2020)

[12] M. Tanabashi et al. (Particle Data Group), Phys. Rev. D 98, 030001 (2018) and 2019 update