



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Inżynierii Metali i Informatyki Przemysłowej

Katedra Informatyki Stosowanej i Modelowania

Praca dyplomowa

Opracowanie algorytmu do poszukiwań lokalnych asymetrii na
rozkładach Dalitza w trzyciałowych rozpadach barionów
powabnych

Developing algorithm for local asymmetries searches in the Dalitz
plot in three-body charm baryon decays

Autor: Dawid Justyna

Kierunek studiów: Informatyka Techniczna

Opiekun pracy: dr hab. Artur Ukleja, prof. AGH

Kraków, 2024

*Serdeczne podziękowania dla mojego promotora,
Pana dr hab. Artura Uklei, prof. AGH
za poświęcony mi czas, przekazaną wiedzę,
a także uwagi dotyczące pisania pracy.
Dzięki Pana wiedzy mogłem
zrozumieć istotę problemu
zwłaszcza w tematyce fizyki cząstek
i ukończyć projekt dyplomowy.*

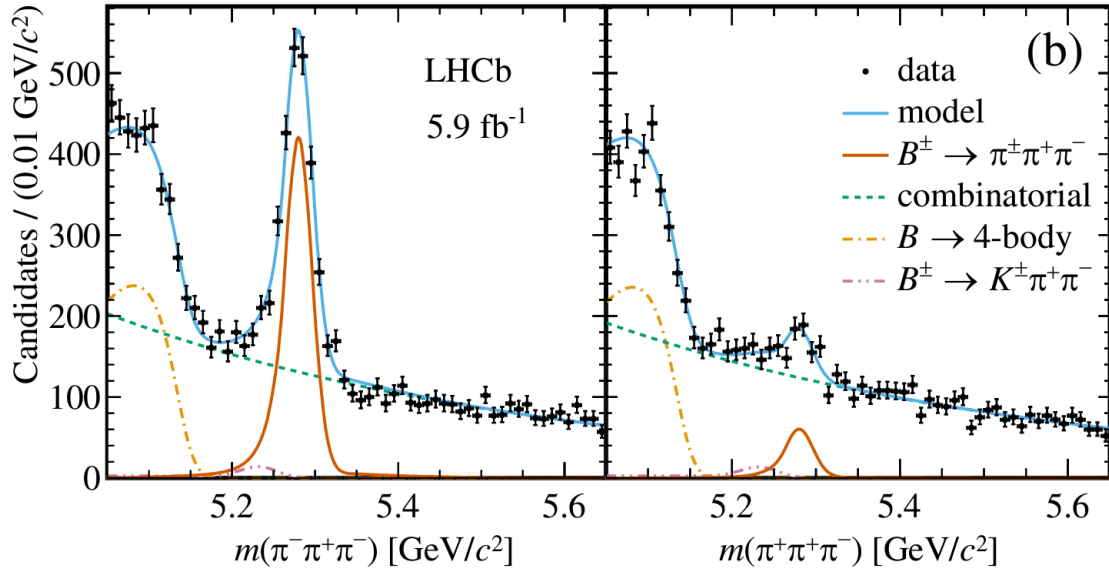
Spis treści

1	Wprowadzenie	4
2	Użyte dane	7
2.1	Rozkłady Dalitza	8
2.2	Próbka bez różnic	9
2.3	Próbka z różnicami	11
3	Zastosowane metody pomiarowe	12
3.1	Metoda S_{CP}	12
3.2	Algorytm Imana-Conovera	14
3.3	Sztuczne sieci neuronowe	21
4	Pomiary różnic między cząstkami i antycząstkami	27
4.1	Metoda S_{CP}	27
4.1.1	Wyniki otrzymane na jednej próbce	27
4.1.2	Wyniki otrzymane na 100 próbkach	31
4.1.3	Czułość metody	32
4.2	Algorytm Imana-Conovera	33
4.2.1	Transformacja danych	33
4.2.2	Wpływ na czułość metody S_{CP}	35
4.3	Sztuczne sieci neuronowe	43
4.3.1	Sposób postępowania	44
4.3.2	Budowa modelu	45
4.3.3	Dobór liczby przedziałów na rozkładach Dalitza	46
4.3.4	Dobór statystycznej wielkości zbiorów treningowego i testowego	47
4.3.5	Dobór parametrów sieci neuronowej	51
4.3.6	Czułość modelu	53
5	Podsumowanie	58
	Literatura	60

1 Wprowadzenie

Zjawisko łamania symetrii między materią, a antymaterią (łamanie parzystości przestrzenno-ładunkowej CP) jest jedną z najmniej poznanych części modelu standardowego. Łamanie parzystości CP jest wynikiem złożenia dwóch przekształceń. Jedno z nich polega na zamianie ładunku (C), np. z dodatniego na ujemny, a drugie polega na inwersji przestrzennej (P), np. zamiana współrzędnych x, y, z na $-x, -y, -z$. Jeżeli jednoczesna transformacja C i P nie powoduje zmian własności układu fizycznego, to oznacza to pełną symetrię między relatywistycznymi cząstkami i antycząstkami. Jeśli natomiast własności układu fizycznego zmieniają się w wyniku tej transformacji, to oznacza to, że prawa fizyki też się zmieniają jeśli zastąpi się cząstkę jej antycząstką. Oznacza to, że obserwowany proces nie jest zwierciadlanym odbiciem początkowej reakcji i w tej sytuacji mówimy, że symetria między materią, a antymaterią jest złamana.

Istnienie niezerowego łamania parzystości CP w przyrodzie zostało potwierdzone w rozpadach mezonów K , B i D w wielu eksperymentach [1–8]. Jednakże wielkość tego łamania jest niewystarczająca do wyjaśnienia obserwowanej we wszechświecie nadwyżki materii nad antymaterią. Dlatego poszukiwane są nowe źródła łamania parzystości CP , jak np. w rozpadach barionów pięknych i powabnych Λ_b^0 , Λ_c^+ i Ξ_c^+ , gdzie jak dotychczas zjawisko to nie zostało potwierdzone [9–13]. Na szczególną uwagę zasługują bariony powabne, gdzie przewidywania łamania parzystości CP są podobnego rzędu jak w przypadku mezonów powabnych i są oczekiwane na poziomie promila, a nawet mniej [14–16]. Wielkość tego efektu jest tu znacząco mniejsza niż wartości mierzone w rozpadach mezonów pięknych B [17, 18], gdzie łamanie parzystości CP wynosi nawet rzędu 75% [19]. Do obserwacji tak dużego efektu wystarczy wizualne porównanie rozkładów mas cząstek i antycząstek, np. przez naniesienie ich na siebie na jednym wykresie lub przedstawienie na dwóch wykresach i identycznej skali. W wyniku łamania parzystości CP na poziomie 75% rozkłady mas cząstek i antycząstek będą o taką wartość różniły się od siebie. Jest to przedstawione na rysunku 1, gdzie maksimum związane z rozpadem antycząstki B^- jest znacząco wyższe niż odpowiadające mu maksimum pochodzące z rozpadu cząstki B^+ , które tylko nieznacznie uwypukla się ponad tło.



Rysunek 1: Rozkłady masy mezonu B rozpadającego się na trzy piony. Na lewym rysunku rozkład masy antycząstki B^- , a na prawym rysunku rozkład masy cząstki B^+ . Dane eksperymentu LHCb są przedstawione w postaci punktów. Linia niebieska to dopasowany model do wszystkich danych. Linia czerwona przedstawia dane po odjęciu tła. Różne wkłady tła są przedstawione w postaci pozostałych linii opisanych na rysunku. Rysunek zaczerpnięty z [19].

Przeciwnie jest to w przypadku cząstek powabnych, gdyż efektów rzędu promila nie da się zobaczyć „gołym” okiem, dlatego zachodzi potrzeba stosowania tu metod czułych na tak małe asymetrie. W niniejszej pracy podjęto się próby zwiększenia czułości standardowo stosowanej metody w eksperymencie LHCb [20] (metoda jest opisana w rozdziale 3.1) oraz zaproponowano i przeprowadzono testy całkowicie nowego podejścia opartego na sieciach neuronowych. Zastosowane metody mogą być użyte do poszukiwania jakichkolwiek różnic między dwoma zbiorami danych, a nie tylko do poszukiwania łamania parzystości CP . Dlatego w pracy skupiono się na opracowaniu narzędzi i ich optymalizacji pod kątem bardzo małych różnic między badanymi próbkami, a nie na źródłach i sposobach łamania CP . Z tego też powodu w pracy nie został opisany eksperyment LHCb, gdyż zaproponowane nowe metody mogą być stosowane nie tylko w każdym eksperymencie do poszukiwania jakichkolwiek różnic, ale także w innych dziedzinach nauk. W przypadku analizy danych eksperymentalnych metody te mogą być ponadto używane do badania efektów detektorowych, czy innych zjawisk fizycznych, które opierają się na wszelkiego rodzaju asymetriach między badanymi rozkładami, jak porównanie danych generowanych z danymi rzeczywistymi. Natomiast jeśli chodzi o inne dziedziny nauk,

to metody mogą być używane np. w medycynie do porównywania zdjęć pacjentów z tomografu komputerowego, jak, np. zdjęć płuc w celu poszukiwania w nich zmian spowodowanych rozrostem raka, czy innych chorób.

W pracy zrealizowano dwa podejścia. Jednym z nich był algorytm Imana-Convera opracowany na początku lat 80-tych poprzedniego wieku. Drugim podejściem było wykorzystanie coraz to szerzej stosowanej sztucznej inteligencji, a dokładniej sztucznych sieci neuronowych. Nie ma obecnie dziedziny, w której sztuczna inteligencja nie odgrywałaby znaczącej roli. Dlatego ciekawym wydaje się być kontrast między dwoma metodami. Z jednej strony algorytm znany bardzo małej liczbie osób, a z drugiej jedna z najbardziej rozwijanych obecnie dziedzin informatyki. Za cel pracy wyznaczono zaimplementowanie narzędzi do wykrywania małych różnic między próbkami, a także porównanie ich z obecnie stosowaną metodą w eksperymencie LHCb.

Niniejsza praca składa się z pięciu rozdziałów, których zakres przedstawia się następująco:

- rozdział 2 zawiera opis o pochodzeniu wykorzystanych danych;
- rozdział 3 zawiera opis używanych metod i algorytmów w projekcie;
- rozdział 4 zawiera opis przygotowywania danych oraz implementacji algorytmów i odczyt otrzymanych wyników;
- rozdział 5 zawiera wnioski i podsumowanie pracy.

2 Użyte dane

Badania przedstawione w niniejszej pracy są oparte na danych generowanych. W celu danych został użyty algorytm napisany na potrzeby eksperymentu LHCb [20], który to został zbudowany między innymi w celu pomiaru łamania parzystości CP . Użyty generator wykorzystuje algorytmy uczenia maszynowego „Tensor Flow” [21]. Nie uwzględnia on oddziaływań cząstek w detektorze, a jedynie generuje sam proces i umożliwia wprowadzenie łamania parzystości CP . Dlatego uzyskane w ten sposób dane są niezależne od jakiegokolwiek układu eksperymentalnego, a otrzymane wyniki będą pomocne każdemu eksperymentowi zajmującemu się pomiarami łamania parzystości CP .

Przy generacji trzyciałowych rozpadów barionów powabnych $\Xi_c^+ \rightarrow pK^-\pi^+$ przyjęto pośrednie stany rezonansowe, jak obserwowane na danych LHCb [12]. Łącznie 18 pośrednich stanów rezonansowych, którymi są: $\Delta(1232)$, $\Delta(1600)$, $\Delta(1620)$, $\Delta(1700)$, K^* , $K^*(1410)$, $K_2^*(1430)$, $K^*(1430)$, $\Delta(1520)$, $\Delta(1600)$, $\Delta(1670)$, $\Delta(1690)$, $\Delta(1710)$, $\Delta(1800)$, $\Delta(1810)$, $\Delta(1820)$, $\Delta(1830)$, $\Delta(1890)$.

Wygenerowane zostały dwa rodzaje danych. Jedne zawierające asymetrię (niżej opisany sposób wprowadzania jej) oraz dane bez asymetrii między cząstkami i antycząstkami będące wynikiem łamania parzystości CP . Oba zestawy zawierały taką samą statystykę danych. Wygenerowano łącznie 3 400 000 rozpadów z barionami powabnymi: 1 700 000 cząstek i 1 700 000 antycząstek. Z tych danych utworzono 100 próbek, na których przeprowadzono badania czułości metod. Przyjęto, że każda z próbek będzie zawierała 100 000 cząstek i tyle samo antycząstek. Taka statystyka danych jest porównywalna do dotychczas zebranej liczby rozpadów $\Xi_c^+ \rightarrow pK^-\pi^+$ w eksperymencie LHCb [13]. Ponieważ łączna statystyka danych potrzebna do utworzenia 100 próbek jest znacząco większa niż wygenerowana oraz biorąc pod uwagę wyjątkowo długi czas generacji nowych zdarzeń, zdecydowano się na użycie techniki statystycznej opartej na repróbkowaniu i w języku angielskim nazywanej bootstrap (były próby przetłumaczenia bootstrap na język polski, ale spełżyły na niczym; próbowano stosować nazwy jak „metody sznurowadłowe”, „metody samowsporne”). Polega to na losowaniu przypadków z dostępnego całościowego zbioru tworząc mniejsze statystycznie próbki danych, jednocześnie zachowując cechy macierzystego zbioru. Do każdej ze 100 próbek wylosowano po 100 000 cząstek i 100 000 antycząstek. Użyto metody prostej próby losowej (z ang. simple random sample) [22], stosując losowanie bez powtórzeń. W pojedynczej próbce nie mogły powtarzać się przypadki, ale między próbkami już tak.

Przy generacji próbek przyjęto zasadę, że jeżeli funkcja falowa opisująca rozpad cząstek jest wyrażona w sposób:

$$Ae^{i\phi}, \quad (1)$$

gdzie A jest amplitudą na rozpad cząstki z fazą ϕ , to funkcja falowa opisująca antycząstki będzie wyrażona w sposób analogiczny:

$$\bar{A}e^{i\bar{\phi}}, \quad (2)$$

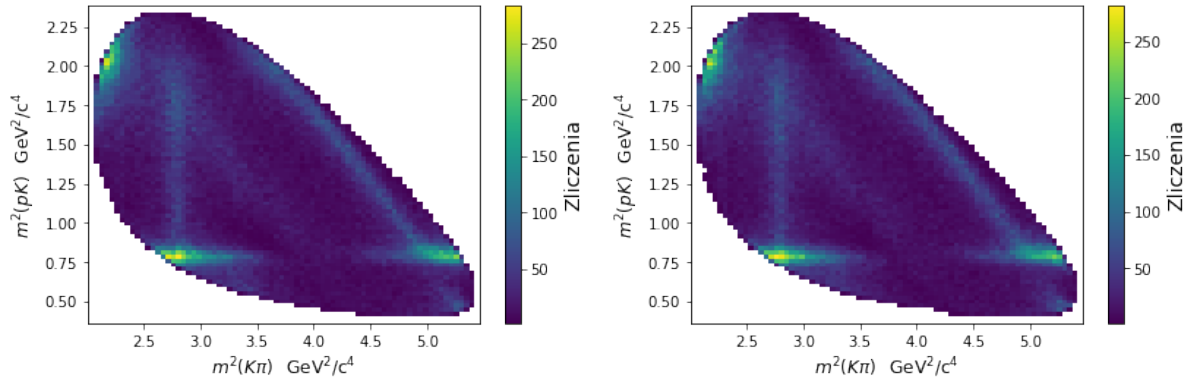
gdzie \bar{A} jest antyamplitudą na rozpad antycząstki z fazą $\bar{\phi}$. W przypadku próbek z brakiem łamania parzystości CP $A = \bar{A}$ i $\phi = \bar{\phi}$. Natomiast w przypadku próbek z łamaniem parzystości CP amplitudy i antyamplitudy i/lub fazy i antyfazy nie są takie same. Do testów wygenerowano próbki z łamaniem parzystości CP (łamanie parzystości CP zamiennie w tej pracy, w celu uproszczenia, jest nazywane asymetrią) 5% w amplitudzie rezonansu $K^*(892)$. Oznacza to, że przy generacji antycząstek antyamplitudę zmieniono o 5%. Jako punkt startowy odpowiadający zerowej asymetrii między cząstkami i antycząstkami, wartości amplitud (i antyamplitud) wszystkich rezonansów przyjęto arbitralnie jako jeden, a fazy (i antyfazy) jako 20° . Uczyniono to dlatego, że w przypadku trzyciałowych rozpadów barionowych nie ma zadowalającego modelu opisującego dane eksperymentalne i w związku z tym nie jest oczywiste jakie przyjąć wartości amplitud i faz poszczególnych stanów rezonansowych. Aktualnie w LHCb są prowadzone analizy celem zbudowania takiego modelu.

2.1 Rozkłady Dalitza

Kinematyka trzyciałowych rozpadów niespolaryzowanych cząstek, takich jak $\Xi_c^+ \rightarrow pK^-\pi^+$ jest opisana przez przestrzeń dwuwymiarową. Może to być jeden z dwóch możliwych rozkładów Dalitza. Na jednym z nich przedstawiony jest na jednej z osi, np. osi y kwadrat masy układu dwóch cząstek stanu końcowego (np. pK^-), a na osi x kwadrat masy układu innej kombinacji dwóch cząstek (np. $K^-\pi^+$). Drugi rozkład Dalitza jest analogiczny i na jednej z osi przedstawiony jest kwadrat masy trzeciej kombinacji dwóch cząstek ($p\pi^+$), a na drugiej osi jeden z kwadratów mas z pierwszego rozkładu Dalitza (np. pK^-). Dlatego przy użyciu wyżej opisanego modelu bezpośrednio wygenerowano kwadraty mas kombinacji dwóch cząstek, które posłużyły do utworzenia rozkładów Dalitza.

2.2 Próbką bez różnic

Przykłady rozkładów Dalitza ze statystyką 100 000 cząstek i 100 000 antycząstek oraz brakiem łamania symetrii CP są przedstawione na rysunku 2. Przyjęte tutaj oznaczenia są stosowane w dalszej części pracy jako $m_1^2 = m^2(K^-\pi^+)$ oraz $m_2^2 = m^2(pK^-)$. Pośrednie stany rezonansowe są widoczne jako zaciemnione obszary w różnych częściach rozkładów Dalitza, które są wynikiem rozpadu danego stanu rezonansowego co prowadzi do zwiększonej produkcji cząstek w danym regionie.



Rysunek 2: Przykładowe rozkłady Dalitza bez asymetrii CP . Rysunek po lewej stronie przedstawia rozkład Dalitza odpowiadający 100 000 cząstek, a po prawej rozkład Dalitza odpowiadający 100 000 antycząstek. Na biało zaznaczone są przedziały, gdzie nie znaleziono żadnych wystąpień. Na osi x przedstawiony kwadrat masy kombinacji cząstek $K\pi$ ($m^2(K\pi)$), na osi y kwadrat masy kombinacji pK ($m^2(pK)$).

Kwadrat masy trzeciej kombinacji pary cząstek ($m^2(p\pi)$) można wyznaczyć korzystając z następującego wzoru:

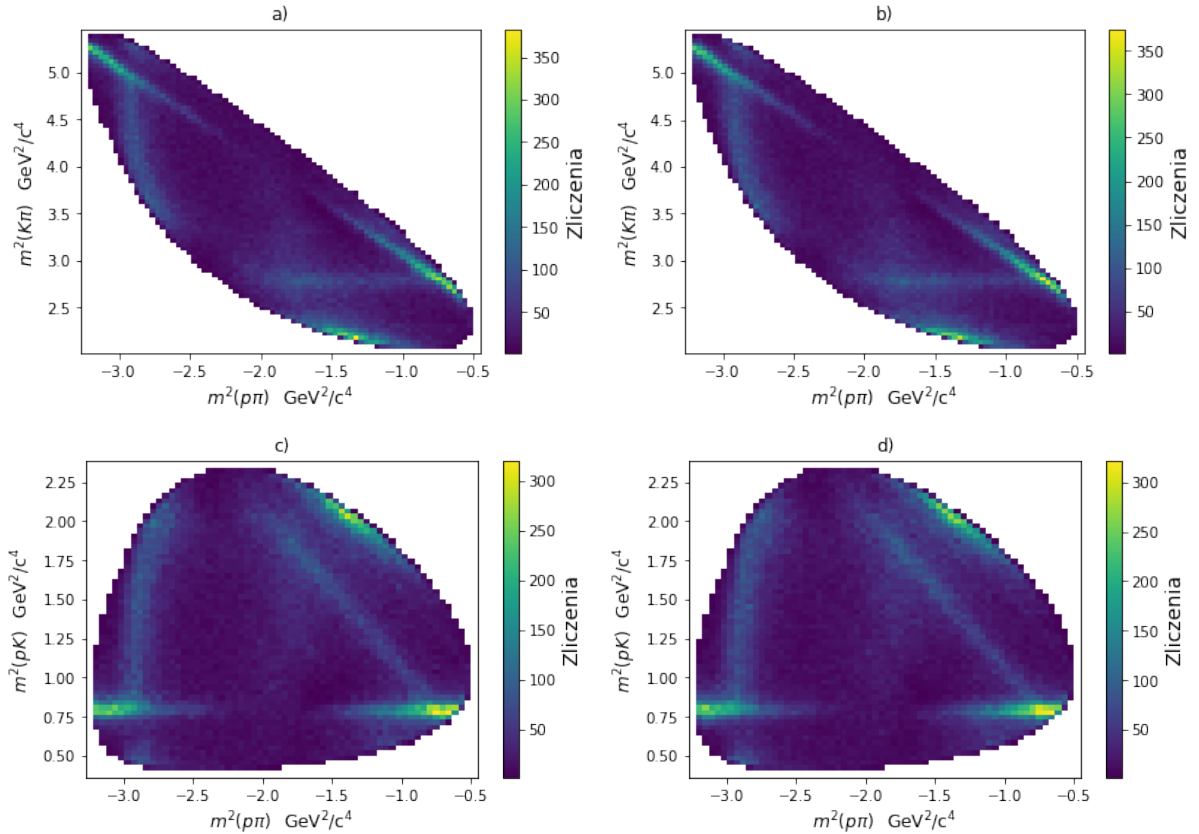
$$m_{pK}^2 + m_{p\pi}^2 + m_{K\pi}^2 = M_{\Xi_c^+}^2 + m_p^2 + m_K^2 + m_\pi^2, \quad (3)$$

gdzie $M_{\Xi_c^+}$ jest masą cząstki macierzystej, w tym przypadku barionu Ξ_c^+ , m_p jest masą protonu, m_K jest masą kaonu, a m_π jest masą pionu. Wartości stałych Ξ_c^+ , m_p , m_K i m_π są podane w tabeli 1.

Tabela 1: Wartości mas wybranych cząstek użytych do obliczeń $m^2(p\pi)$.

Cząstka	Masa w jednostkach GeV/c^2
Ξ_c^+	1,315
p	0,938
K^-	0,493
π^+	0,013

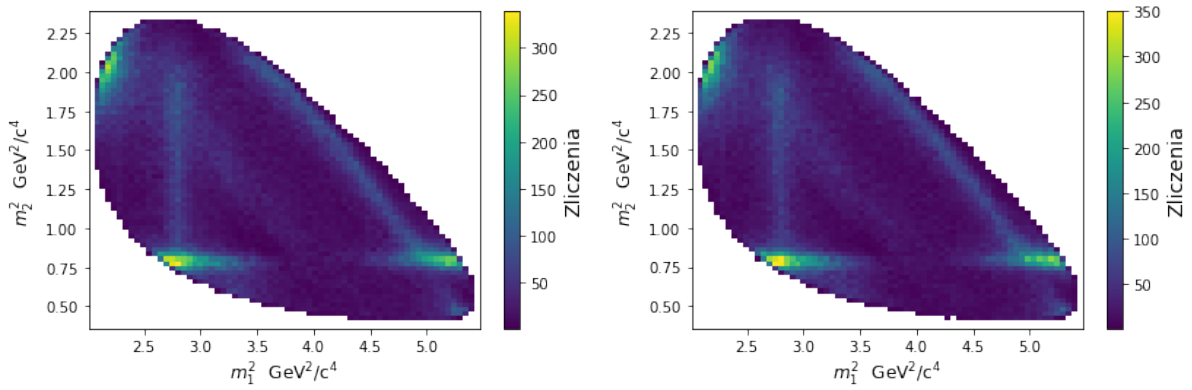
Korzystając ze wzoru 3 i danych z tabeli 1 obliczono $m^2(p\pi)$. Rozkłady Dalitza, gdzie na jednej osi użyta jest wartość $m^2(p\pi)$ są przedstawione na rysunku 3. Widoczne poprzeczne (diagonalne) zacienienia są odbiciem trzeciej kombinacji pary cząstek. Dlatego w celu poszukiwania sygnałów wynikających z łamania parzystości CP wystarczy badać tylko jeden typ z przedstawionych rozkładów Dalitza (np. m_1^2 i m_2^2) i poszukiwać różnic między rozkładami odpowiadającymi cząstkom i antycząstkom.



Rysunek 3: Rozkłady Dalitza utworzone z $m^2(p\pi)$ (oś x) i $m^2(K\pi)$ (oś y): cząstki (a) i antycząstki (b) oraz utworzone z $m^2(p\pi)$ (oś x) i $m^2(pK)$ (oś y): cząstki (c) i antycząstki (d). Użyto 100 000 cząstek i 100 000 antycząstek.

2.3 Próbką z różnicami

Na rysunku 4 przedstawiono rozkłady Dalitza otrzymane na danych z asymetrią w konfiguracji m_1^2, m_2^2 . Te rysunki pokazują, że nie da się zauważyć „gołym” okiem różnic spowodowanych łamaniem parzystości CP między rozkładami cząstek i antycząstek, jak również między rozkładami bez asymetrii, które zostały przedstawione na rysunku 2. Dlatego konieczne jest poszukiwanie metod, które będą w stanie odpowiedzieć na pytanie, czy porównywane rozkłady są tożsame, czy też statystycznie różne od siebie, czyli metod znajdowania różnic między rozkładami (jak różnic spowodowanych asymetrią CP). W tym celu zaimplementowano algorytm Imana-Conovera do poprawy czułości standardowej metody stosowanej w eksperymencie LHCb oraz zaproponowano użycie sztucznych sieci neuronowych.



Rysunek 4: Rozkłady Dalitza otrzymane na danych z asymetrią 5%. Rozkład Dalitza odpowiadający 100 000 cząstek po lewej stronie oraz rozkład Dalitza odpowiadający 100 000 antycząstek po prawej stronie.

Podsumowując, do badań przedstawionych w tej pracy użyto 100 próbek z wygenerowanymi rozkładami Dalitza (m_1^2, m_2^2) trzciałowych rozpadów barionów powabnych $\Xi_c^+ \rightarrow pK^-\pi^+$ bez łamania parzystości CP (w skrócie nazywanych danymi bez asymetrii) i 100 analogicznych próbek z łamaniem parzystości CP uwzględnionej w amplitudach rezonansu K^* na poziomie 5% (w skrócie nazywanych danymi z asymetrią). W każdej próbce użyto takiej samej liczby cząstek i antycząstek po 100 000. Taka statystyka jest porównywalna do zebranych danych doświadczalnych w eksperymencie LHCb [13].

3 Zastosowane metody pomiarowe

W eksperymencie LHCb łamanie parzystości CP w rozpadach barionów powabnych było badane z użyciem kilku metod. Porównanie tych metod jest opisane w monografii [13]. Żadna z tych metod nie potwierdziła istnienia łamania parzystości CP w rozpadach jakiegokolwiek barionu, zarówno pięknego, jak i powabnego. Trudno zatem jest odpowiedzieć na pytanie, czy przyczyną tego jest zbyt mała statystyka danych, czy też użyte metody nie były dostatecznie czułe na oczekiwane łamanie parzystości CP (rzędu promili w rozpadach cząstek powabnych). Dlatego w niniejszej pracy skupiono się na próbach zwiększenia czułości badań przy tej samej statystyce danych. Zaproponowano tu dwa nowe podejścia nierozpatrywane dotychczas: algorytm Imana-Conovera i sieci neuronowe. Czułości nowych podejść są badane w odniesieniu do standardowej metody poszukiwań lokalnych różnic między cząstkami i antycząstkami, tzw. metody S_{CP} . Algorytmy oraz metody zaimplementowano w języku Python 3.

3.1 Metoda S_{CP}

Metoda S_{CP} wykorzystywana jest do poszukiwania lokalnych asymetrii. Bazuje ona na podziale rozkładu Dalitza na przedziały, które następnie są porównywane ze swoim odpowiednikiem cząstek i antycząstek. Przedziały mogą być różnej liczby i tych samych lub różnych rozmiarów. W tej pracy zastosowano podział rozkładów Dalitza na przedziały jednakowych rozmiarów. Po podziale rozkładów Dalitza cząstek i antycząstek na jednakowe przedziały pod względem ich liczby i rozmiarów, w celu porównania cząstek i antycząstek, przedział po przedziale obliczana jest wartość S_{CP}^i zdefiniowana w następujący sposób:

$$S_{CP}^i = \frac{n_+^i - \alpha n_-^i}{\sqrt{\alpha(n_+^i + n_-^i)}}, \quad (4)$$

gdzie: i oznacza numer przedziału, n_+^i jest liczba cząstek w i -tym przedziale, n_-^i jest liczba antycząstek w tym samym przedziale, a α jest współczynnikiem kasującym globalne asymetrie w przypadku występowania różnic między zsumowaną liczbą cząstek $N^+ = \sum_{i=1}^{N^+} n_+^i$ i antycząstek $N^- = \sum_{i=1}^{N^-} n_-^i$. W eksperymencie nie zawsze całkowite liczby cząstek i antycząstek są takie same. Różne fizyczne efekty mogą prowadzić do produkcji różnej liczby cząstek i antycząstek. W przypadku akceleratora LHC, który dostarcza wiązki protonów do eksperymentu LHCb, zderzenia proton-proton nie są symetryczne ładunkowo (symetryczne byłyby gdyby

były zderzane protony z antyprotonami). Prowadzi to do różnic w produkcji liczby cząstek i antycząstek podczas zderzeń. Jest to nazywane asymetrią produkcji, która jest mierzona na poziomie kilku procent. Dlatego w celu pozbycia się tego typu efektów służy współczynnik α , a mierzona wartość S_{CP}^i w każdym przedziale będzie czuła tylko na lokalne efekty wynikające z różnic między cząstkami i antycząstkami. Te różnice mogą być spowodowane, między innymi, łamaniem symetrii CP . Ponadto tak zdefiniowana wartość S_{CP}^i (pierwiastek w mianowniku) jest znaczącością różnicy między cząstkami i antycząstkami, a nie bezwymiarową ich różnicą (gdyby nie było pierwiastka w mianowniku).

W przypadku wygenerowanych próbek danych $\alpha = 1$, gdyż liczba cząstek i antycząstek w każdej próbie jest taka sama i wynosi po 100 000. W takiej sytuacji wzór 4 można uprościć do następującej postaci:

$$S_{CP}^i = \frac{n_+^i - n_-^i}{\sqrt{n_+^i + n_-^i}}. \quad (5)$$

W celu pozbycia się czułości na fluktuacje statystyczne brane pod uwagę są tylko te przedziały, w których statystyka cząstek jak i antycząstek jest większa od dziesięć. Przedziały nie spełniające tego warunku są odrzucane z pomiarów. Tak zdefiniowane S_{CP}^i jest nie tylko znaczącością różnicy między cząstkami i antycząstkami a dodatkowo suma kwadratów tych wartości po wszystkich analizowanych przedziałach jest testem χ^2 . Można to zapisać następującym wzorem:

$$\chi^2 \equiv \sum (S_{CP}^i)^2. \quad (6)$$

Zdefiniowane testu χ^2 pozwala na wyznaczenie prawdopodobieństwa (w dalszej części pracy używany jest zapis p-value) zgodności z hipotezą zerową. Jako hipotezę zerową przyjęto zgodność statystyczną rozkładów Dalitza cząstek i antycząstek, czyli brak asymetrii. W takiej sytuacji rozkład wartości S_{CP}^i będzie rozkładem normalnym o średniej $\mu = 0$ i odchyleniu standardowym $\sigma = 1$. W przypadku niezerowej asymetrii, czyli niezerowych statystycznych różnic między cząstkami i antycząstkami rozkład wartości S_{CP}^i będzie różny od rozkładu normalnego i wartość p-value będzie odpowiednio mała. W przypadku braku zgodności z hipotezą zerową, czyli w przypadku różnic między cząstkami i antycząstkami przekraczającymi pięć standardowych odchyłeń mierzone p-value będzie mniejsze niż $3 \cdot 10^{-7}$.

3.2 Algorytm Imana-Conovera

W 1982 roku amerykańscy matematycy Ronald L. Iman i William Jay Conover opracowali metodę do radzenia sobie z zależnymi zmiennymi losowymi [23]. Wykorzystali oni rangi zmiennych do tworzenia korelacji. Rangowaniem nazywa się uporządkowanie rosnąco lub malejąco wartości danej zmiennej i na tej podstawie przypisaniu im numerów kolejności występowania. W algorytmice sprowadza się to do posortowania danej zmiennej i przypisaniu jej wartościom indeksów występowania w uporządkowanym już zbiorze. Metoda Imana-Conovera opiera się na ustalaniu korelacji za pomocą macierzy Choleskiego [24], w której poddaje się macierz korelacji pod tak zwaną transformację Choleskiego, a następnie wynik jej pomnożony zostaje przez wejściowy zestaw rang danych. Macierz Choleskiego została opracowana przez francuskiego matematyka André-Louis Cholesky i sprowadza się do dość prostego wzoru, polegającego na przedstawieniu danej macierzy za pomocą iloczynu innej macierzy i jej transpozycji [25], co można zapisać w następujący sposób:

$$X = L * L^T, \quad (7)$$

gdzie X jest macierzą poddawaną transformacji, L jest macierzą Choleskiego, a L^T jest transpozycją macierzy L .

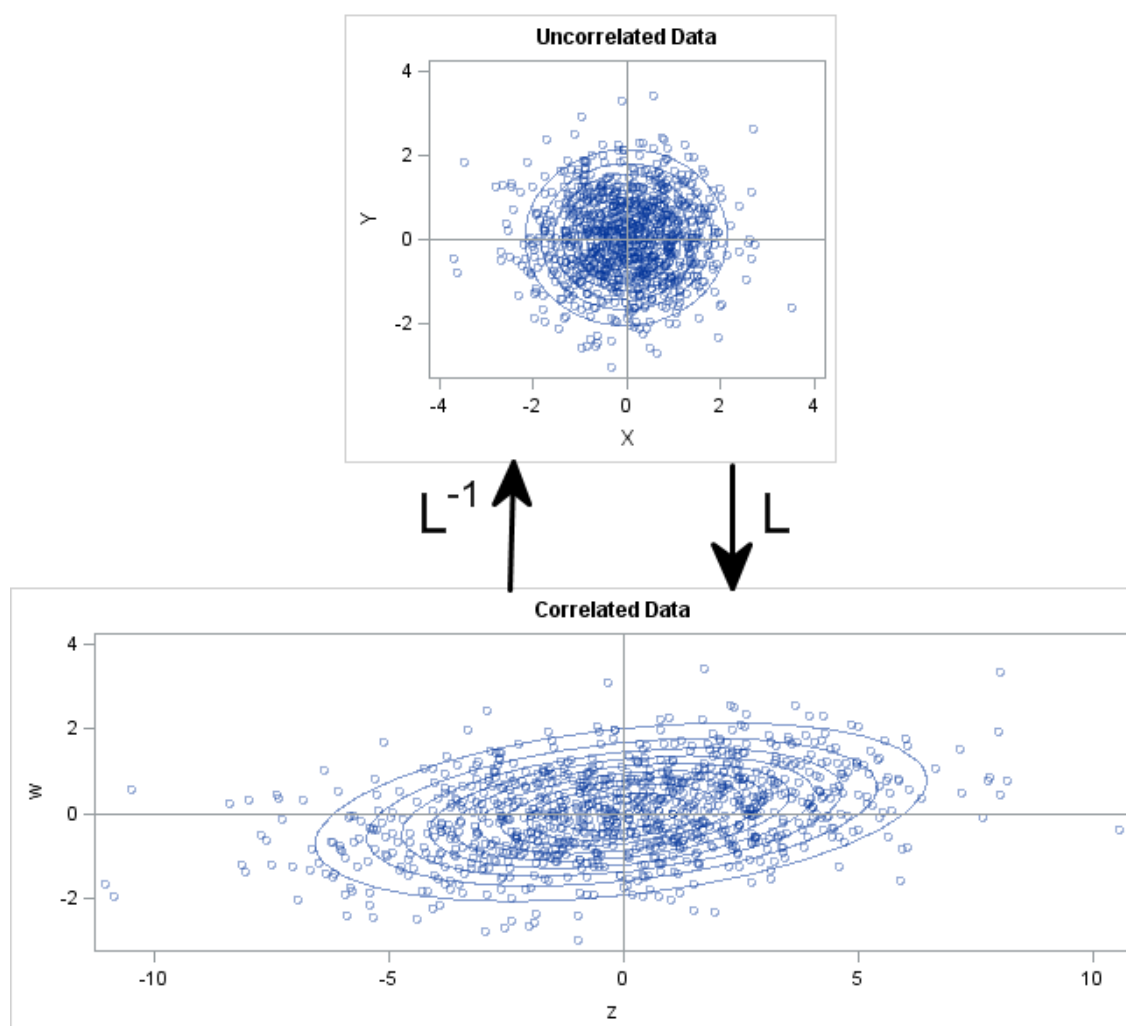
Wyznaczenie macierzy L z równania 7 można przedstawić za pomocą następującej formuły [25]:

$$\begin{cases} L_{i,j} = \sqrt{X_{i,j} - \sum_{k=1}^{j-1} L_{i,k}^2}, & \text{gdzie } i = j, \\ L_{i,j} = \frac{1}{L_{i,j}}(X_{i,j} - \sum_{k=1}^{j-1} L_{i,k} * L_{j,k}), & \text{gdzie } i \neq j, \end{cases} \quad (8)$$

gdzie i jest numerem wiersza, j jest numerem kolumny, n jest liczbą wierszy, $j = 1, 2, \dots, n$, $i = j, j + 1, \dots, n$.

Macierz Choleskiego wykorzystywana jest również do dekorelacji zmiennych. Postępowanie jest podobne, jednak obliczana jest wtedy macierz odwrotna do macierzy Choleskiego. Na rysunku 5 przedstawiono jak w praktyce zmieniają się rozkłady przykładowych danych w przypadku dwóch zmiennych X i Y . Obliczana jest macierz L z oczekiwanej macierzy korelacji, następnie macierz L jest mnożona przez dane wejściowe zapisane w postaci macierzy. Otrzymana w wyniku tej operacji macierz jest danymi z określoną wcześniej nową korelacją. Transformacja Choleskiego jest odwracalna, co można też zauważyć na strzałkach z rysun-

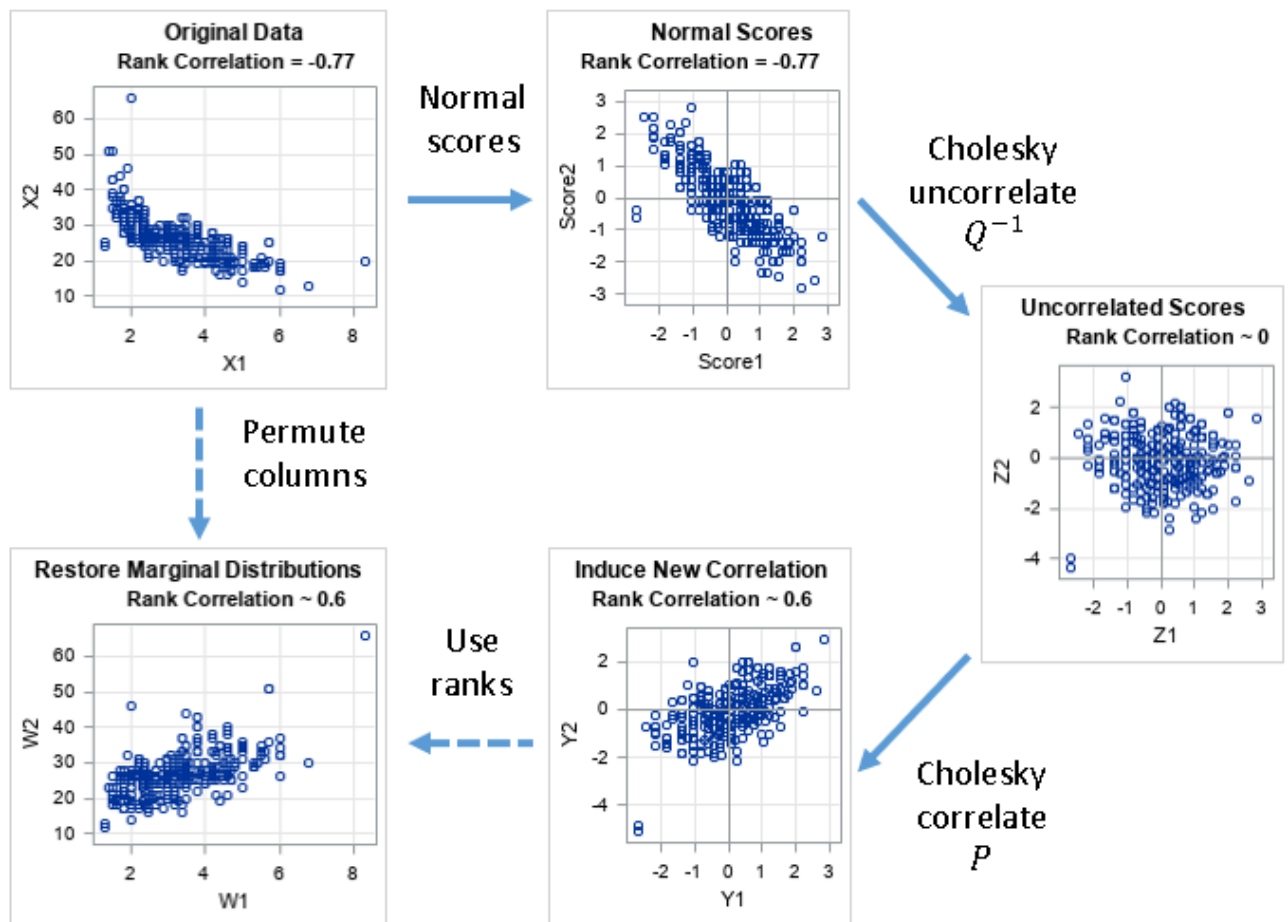
ku 5, gdzie pomnożenie przez macierz L^{-1} wraca do danych wejściowych. Jest to też przykład wspomnianej wcześniej możliwości dekorelacji zmiennych.



Rysunek 5: Korelowanie przykładowych dwóch zmiennych X i Y z wykorzystaniem macierzy Choleskiego. Na rysunku u góry wejściowe dane. Na dole dane otrzymane po nałożonej korelacji (pomnożeniu danych wejściowych przez macierz Choleskiego). Rysunek zaczerpnięty z [24].

Transformacja Imana-Conovera koreluje rangi, a nie konkretne dane. Niesie to za sobą zarówno pozytywne jak i negatywne konsekwencje. Zaletami takiego rozwiązania jest to, że jest ono stosunkowo proste, zakłada podobne kroki co transformacja Choleskiego wykonując je na innych danych. Korelacja rang jest często zbliżona do korelacji liniowej, co również może wpłynąć pozytywnie na wyniki. Mogą wystąpić wyjątki, np. jeśli liczba wartości w zbiorze danych

jest mała, to liczba możliwych permutacji jest wtedy również mała i mogą pojawić się problemy ze znalezieniem odpowiedniej kombinacji. W sytuacji, gdy takie problemy nie występują transformacja Imana-Conovera pozwala bez zmiany rozkładów pojedynczych zmiennych ustawić ich korelacje wraz z innymi zmiennymi zależnymi. Schemat postępowania przedstawiono na rysunku 6.



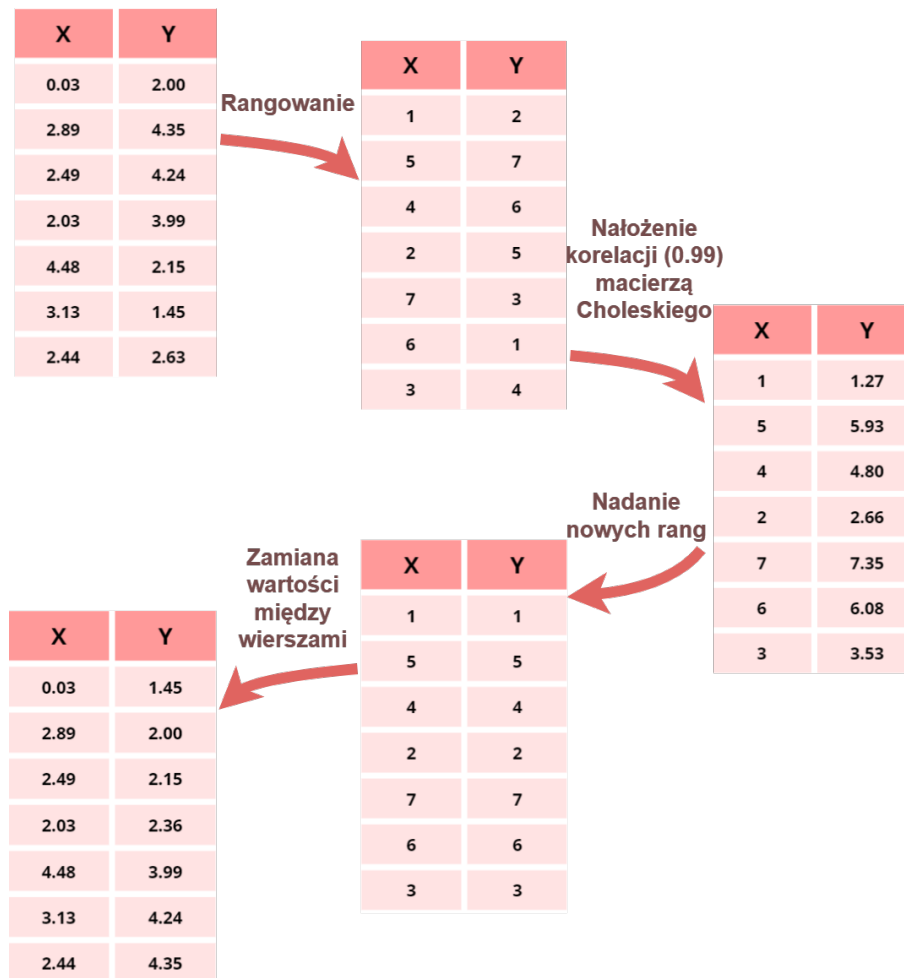
Rysunek 6: Graficzne przedstawienie działania algorytmu Imana-Conovera. Opis podany w tekście. Rysunek zaczerpnięty z [26].

Poszczególne kroki, które należy wykonać w implementacji algorytmu Imana-Conovera można opisać w następujący sposób:

1. normalizacja zmiennych $X_1 \rightarrow \text{Score}_1$, $X_2 \rightarrow \text{Score}_2$, zmienne Score nie zmieniają korelacji wyjściowej;
2. dekorelacja zmiennych (jeśli jakkolwiek korelacja istnieje) odwrotną macierzą Choleskiego: $\text{Score}_1 \rightarrow Z_1$, $\text{Score}_2 \rightarrow Z_2$;
3. nadanie rang zmiennym, czyli ponumerowanie ich miejsc występowania w posortowanym zbiorze w kierunku rosnącym;
4. znormalizowanie rang, a następnie pomnożenie ich przez macierz Choleskiego przy określonych wartościach macierzy korelacji: $Z_1 \rightarrow Y_1$, $Z_2 \rightarrow Y_2$;
5. nadanie rang nowym zmiennym, powstałym po transformacji Choleskiego;
6. zamiana wartości danych wejściowych między wierszami zgodnie z rangami otrzymanymi po transformacji: $Y_1 \rightarrow W_1$, $Y_2 \rightarrow W_2$.

W opisany powyżej sposób nałożona zostaje korelacja nie zmieniając ani jednej wartości otrzymanej na wejściu a permutując jedynie ich miejsce wystąpienia.

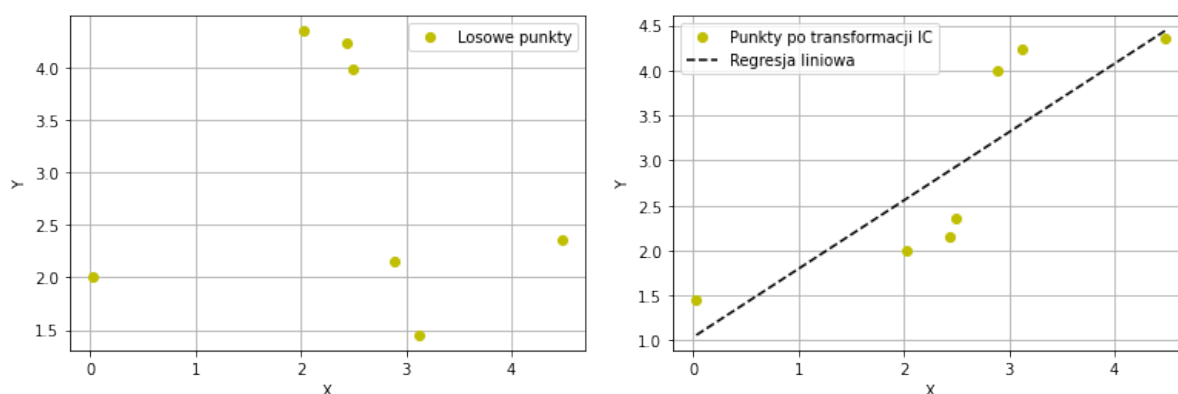
Na rysunku 7 pokazano schemat graficzny jak przebiegają poszczególne kroki transformacji. Schemat wykonany został przy małej próbie dwóch zmiennych z siedmioma przypadkami losowanymi z zakresu od 0 do 5, aby można było dokładnie prześledzić co dzieje się po każdej operacji.



Rysunek 7: Schemat działania algorytmu Imana-Conovera na przykładzie dwóch zmiennych X i Y z siedmioma przypadkami losowanymi z zakresu od 0 do 5. Nałożona korelacja wynosi 0,99.

Na rysunku 7 pierwsza kolumna nie ulega zmianie i w takiej samej kolejności została podana na wejściu oraz na wyjściu. Każda kolejna kolumna (w tym przypadku została jedna kolumna Y) zostaje przemnożona przez odpowiedni wektor pochodzący z macierzy Choleskiego i na tej podstawie dobierane są rangi. W tym przypadku ustawiono korelację na 0,99, czyli niemalże zależność liniową. Z tego też powodu po kroku nadania nowych rang te same kolejności występują w kolumnie X i Y. Pokazuje to, że najmniejsza wartość w zbiorze X odpowiada najmniejszej wartości w zbiorze Y itd., co w przypadku korelacji liniowej jest pożądanym zjawiskiem. Przy innych wartościach korelacji ten wynik nie byłby tak oczywisty. Kolejną rzeczą wartą uwagi jest pierwsza i ostatnia tabela. Wszystkie wartości na wejściu są również na wyjściu. Sprawia to, że rozkłady pojedynczej zmiennej nie ulegają żadnym zmianom. Na

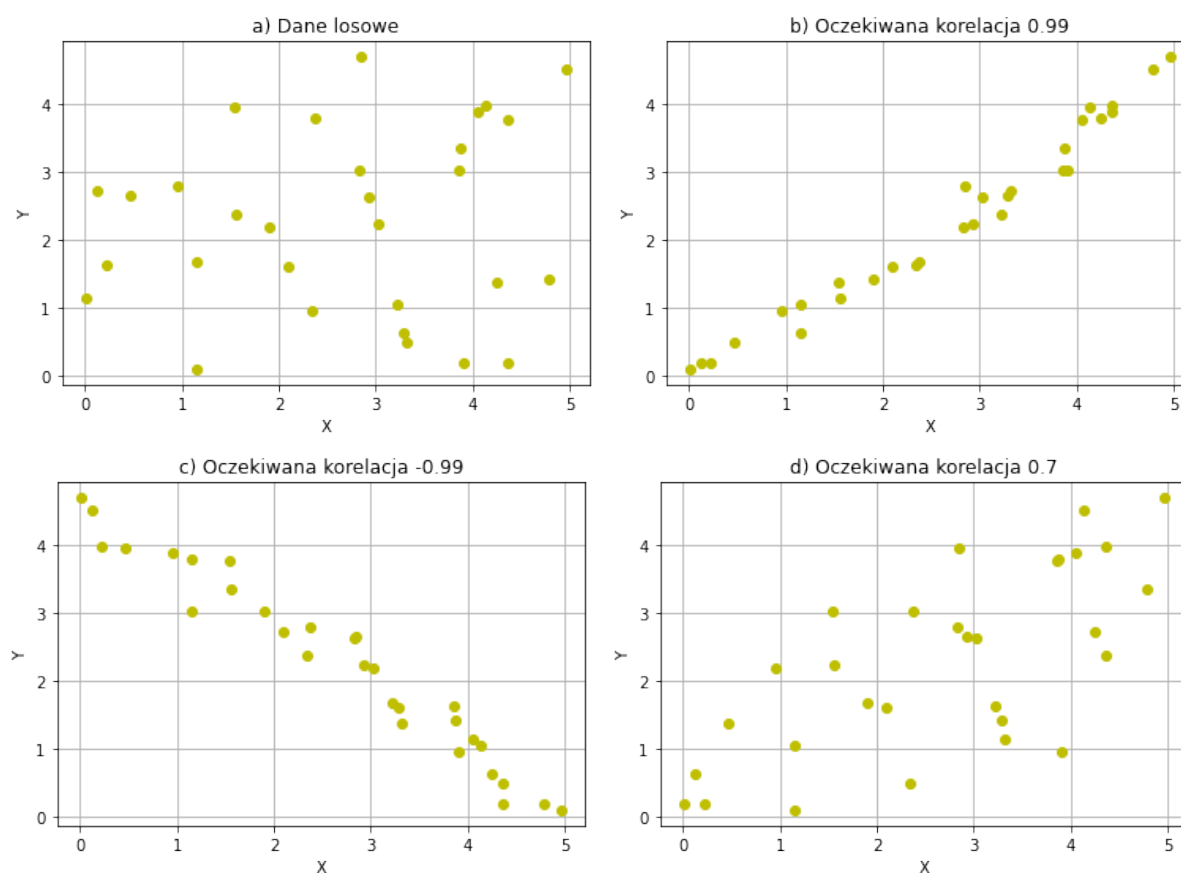
rysunku 8 przedstawiono dane przed transformacją, gdy korelacja wynosiła 0 i po transformacji, gdy oczekiwana korelacja wynosiła 0,99. Do danych po transformacji Imana-Conovera została dopasowana funkcja liniowa o równaniu $Y=0,76 \cdot X+1,04$ (wartości 0,76 i 1,04 zostały otrzymane z dopasowania) w celu zobrazowania jakie odchylenia może powodować metoda. Przy korelacji 0,99 oczekiwanym wynikiem byłoby nałożenie prostej na dane. Widać jednak, że tak nie jest. Korelacja tych danych wyniosła 0,84. Zarówno ta jak i kolejne pojawiające się w tej pracy obliczone wartości korelacji są liczone ze wzoru na współczynnik korelacji Pearsona. Otrzymany wynik odbiega od oczekiwanego o 0,15. Jest to jednak spowodowane małą próbą danych zastosowaną na potrzeby wyjaśnienia algorytmu.



Rysunek 8: Dane X i Y z rysunku 7 przed (lewy wykres) i po (prawy wykres) transformacji Imana-Conovera. Na prawym rysunku do danych dopasowano linię prostą $Y=0,76 \cdot X+1,04$ (wartości 0,76 i 1,04 otrzymane z najlepszego dopasowania).

Na rysunku 9 przedstawiono jak algorytm radzi sobie z analogicznymi danymi jak dotychczas, ale różnymi oczekiwanymi macierzami korelacji. Tym razem zwiększono statystykę próbki z siedmiu do 30. Wyniki działania transformacji Imana-Conovera zostały przedstawione w tabeli 2. Wyniki te pokazują jak dokładna może być metoda Imana-Conovera jeśli próba jest odpowiednio duża (mowa tu tylko o 30 przypadkach). Obserwowane zmiany są nie większe niż 2%. W przypadku analizowanych danych bez asymetrii i z asymetrią w każdej próbce jest 200 000 cząstek i antycząstek łącznie, czyli znacząco więcej niż 30 zliczeń. W tej sytuacji zmiany w wyniku użycia metody Imana-Conovera będą znacząco mniejsze i nie powinny przekraczać 0,03%.

W powyższym teście użyto danych losowo wygenerowanych z przedziału od 0 do 5. W przypadku, gdyby jedna z wartości miała wartość znacząco różniącą się, jak np. 20, to algorytm nie pozwoliłby na zmodyfikowanie takiej wartości i uzyskanie korelacji w okolicach 0,99 byłoby niemożliwe.



Rysunek 9: Zastosowanie algorytmu Imana-Conovera na danych X i Y z 30 przypadkami (a), ale w różnych sytuacjach oczekiwanej korelacji: 0,99 (b), -0,99 (c), 0,70 (d).

Tabela 2: Porównanie wyników transformacji Imana-Conovera na danych z 30 przypadkami.

Oczekiwana korelacja	Otrzymana korelacja po transformacji
0,99	0,97
-0,99	-0,99
0,70	0,68

Transformacja Imana-Conovera nie jest popularną i rozpowszechnianą metodą. Jest to spowodowane tym, że rzadko się zdarza, aby dokładna wartość pojedynczej zmiennej stanowiła cenną informację. Najczęściej cały zestaw danych i występujące w nim zależności są pożądane. Zdarzają się jednak sytuacje, gdzie algorytm ten ma swoje zastosowanie między innymi w symulacjach modeli ryzyka systemowego w bankach [27], czy symulacjach do radioterapii [25].

W przypadku łamania parzystości CP różnice między rozkładami Dalitza powabnych cząstek i antycząstek są oczekiwane jako bardzo małe i w związku z tym trudne do wykrycia. Dlatego w niniejszej pracy użyto transformacji Imana-Conovera w celu sprawdzenia, czy algorytm ten może poprawić czułość metody S_{CP} . Zaletą użycia tego algorytmu jest to, że nie utraci się żadnej informacji, gdyż dane wejściowe nie zostaną zmienione, a jedynie inaczej rozłożone. Dlatego ustawienie korelacji spowoduje tylko inne rozłożenie cząstek i antycząstek na rozkładach Dalitza, jak przesypywanie się cząstek i antycząstek z jednego regionu do innego. To z kolei potencjalnie może wpłynąć na zwiększenie liczby przedziałów w metodzie S_{CP} i tym samym daje nadzieję na zwiększenie czułości tej metody. Można to wyobrazić sobie w następujący sposób: w przypadku kiedy różnice między cząstkami i antycząstkami występują, np. tylko w czterech przedziałach na 20 w metodzie S_{CP} , to po transformacji Imana-Conovera proporcja liczby przedziałów z różnicami między cząstkami i antycząstkami w odniesieniu do całkowitej liczby przedziałów może wzrosnąć. To znaczy, że różnice, które były tylko w czterech przedziałach na 20 rozłożą się na więcej przedziałów. Statystyka w tych przedziałach się zmniejszy, ale można ją utrzymywać na poziomie powyżej fluktuacji statystycznych. Większa liczba przedziałów z różnicami między cząstkami i antycząstkami w porównaniu do całkowitej liczby przedziałów w teście χ^2 oznacza mniejsze p-value w odniesieniu do mierzonego metodą S_{CP} bez transformacji Imana-Conovera. W ten sposób zwiększa to czułość metody S_{CP} przy tej samej statystyce danych.

3.3 Sztuczne sieci neuronowe

Sztuczna inteligencja znajduje coraz szersze zastosowanie w codziennym życiu, a także w rozwoju nauki. Stale rozwijane są nowe metody lub wykorzystywane są starsze, które dzięki rozwojowi technologii są w stanie osiągać zadowalające wyniki. Można wyróżnić niektóre z podziałów algorytmów sztucznej inteligencji, między innymi, na rodzaj ich uczenia:

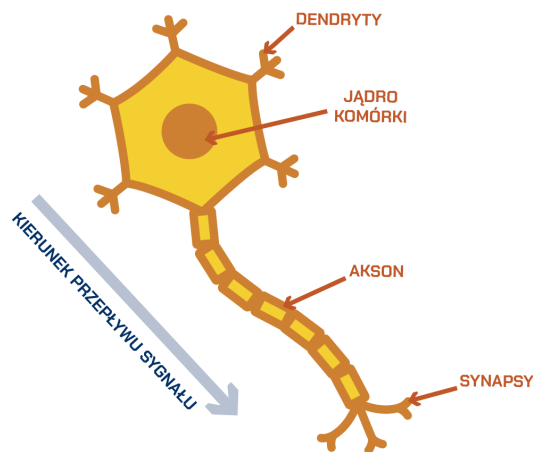
- **Uczenie nadzorowane** - polega na tym, że znany jest cel, do którego ma dążyć model i próbuje się na jego podstawie osiągnąć jak najlepsze wyniki. Dodatkowo można te

algorytmy rozdzielić na dwie grupy:

1. **Klasyfikatory** - dana jest skończona liczba cech i algorytm na wyjściu oblicza prawdopodobieństwo przynależności do jednej z klas. Efektywność takich algorytmów mierzy się na podstawie skuteczności trafionych przewidywań.
 2. **Regresory** - algorytm przewiduje wartość liczbową. Efektywność w tym przypadku mierzy się metodami, które obliczają różnicę między wartością przewidywaną a rzeczywistą, np. w fizyce cząstek elementarnych przewiduje się energię i kierunek lotu mionu pochodzącego z promieniowania kosmicznego.
- **Uczenie nienadzorowane** - charakteryzuje się tym, że nie są znane wyniki, które otrzymywane są na wejściu i nie można nauczyć modelu, jak wygląda to, co ma znaleźć. Celem takich algorytmów jest przypisywanie do poszczególnych zbiorów i szukanie różnic między elementami. Przykładem takich algorytmów jest klasteryzacja, która polega na pogrupowaniu danych w klastry w zależności od ich cech pod względem podobieństwa.
 - **Uczenie przez wzmacnianie** - polega na poruszaniu się tak zwanego agenta. On na podstawie „kar” i „nagród” (przez odpowiednie przypisywanie wag) decyduje czy podąża w kierunku bliskim celu, czy też powinien zmienić kierunek.

W niniejszej pracy podjęto próbę zastosowania uczenia maszynowego do poszukiwania różnic między cząstkami i antycząstkami. Dlatego badania te rozpoczęto od użycia dobrze znanych i stosowanych od wielu lat sieci neuronowych opartych na perceptronie. Testowano tutaj dwa podejścia: klasyczne i przeciwstawne jemu. Klasyczne sieci neuronowe są budowane z wielu neuronów i jednej lub dwóch warstw ukrytych. Natomiast przeciwstawne temu podejście polega na budowie bardzo głębokiej sieci neuronowej, tzw. wielowarstwowej MLP (Multilayer Perceptron) i kilku neuronów w każdej warstwie.

Sieci neuronowe stanowią odwzorowanie ludzkiego układu nerwowego, który składa się z pojedynczych neuronów, gdzie każdy z nich odpowiada za określone procesy.

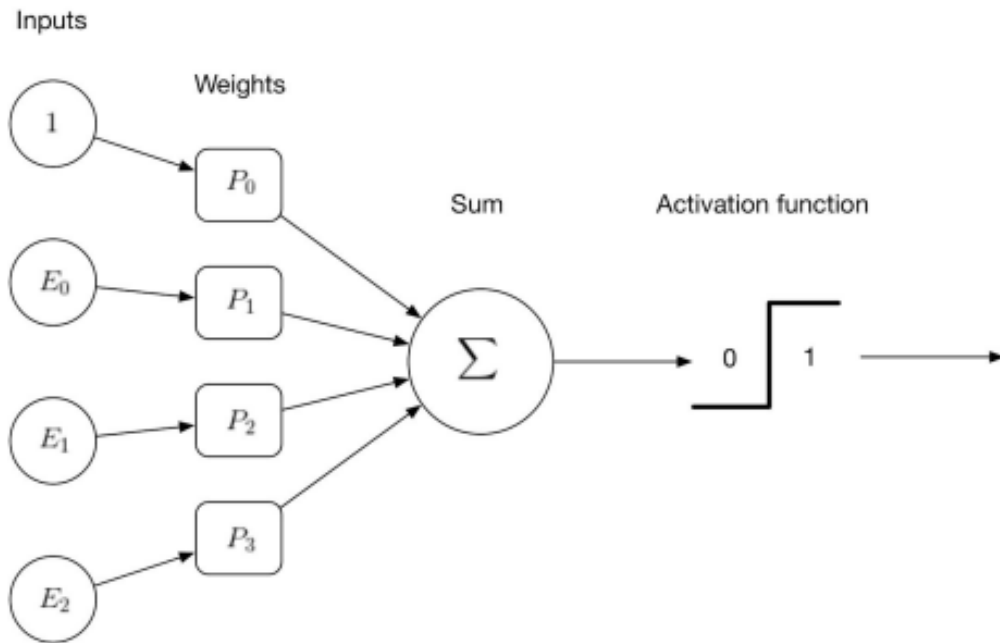


Rysunek 10: Model biologicznego neuronu. Rysunek zaczerpnięty z [28].

Na rysunku 10 przedstawiony jest model biologicznego neuronu. Poszczególne jego elementy odpowiadają za następujące funkcje:

- dendryty – otrzymują sygnały ze środowiska;
- jądro komórkowe – przechowuje informację genetyczną i powiela ją;
- akson – przekazuje sygnał wejściowy dalej;
- synapsy – łączą neuron z innym neuronem przekazując dalej sygnał.

Mimo że sieci neuronowe mają coraz szersze zastosowanie, to postęp wynika głównie z rozwoju technologii, a same metody powstały dużo wcześniej. To już w latach 50. poprzedniego wieku powstała koncepcja sztucznego neuronu [29]. Początkowo opierała się ona na prostym schemacie o wielu wejściach i jednym wyjściu logicznym zwracającym 0 lub 1. Schemat działania sztucznego neuronu (nazywany też perceptronem) jest przedstawiony na rysunku 11.



Rysunek 11: Prosty model sztucznej sieci neuronowej. Rysunek zaczerpnięty z [30].

Sztuczna sieć neuronowa posiada z góry zdefiniowaną liczbę wejść. Na rysunku 11 przedstawiono cztery wejścia oznaczone jako: 1, E_0 , E_1 , E_2 . Każdemu wejściu przypisywana jest waga (P_0 , P_1 , P_2 , P_3), która determinuje jak istotny jest każdy z neuronów. Najczęściej startowe wagi są przypisywane w sposób losowy, a w kolejnych iteracjach są one zmieniane w przyjęty z góry sposób, tak aby jak najlepiej dopasować model do rzeczywistych wyników. Wartość na każdym wejściu jest mnożona przez określoną wagę, które następnie są sumowane. Można to opisać następującym wzorem:

$$Y = P_0 + \sum E_i * P_{i+1}, \quad (9)$$

gdzie $i = 0, 1, 2$.

Poszczególne neurony (E_i) są mnożone przez ich wagi (P_{i+1}) z wyjątkiem wagi P_0 . Wejście, które do niej prowadzi ma zawsze wartość 1. Jest to tak zwany bias. Nie zależy on od żadnego z neuronów, a jego wartość pozwala na przesuwanie funkcji aktywacji (opisana jest w dalszej części pracy).

Po przypisaniu i zsumowaniu wag określana jest funkcja aktywacji na podstawie obliczonej wartości Y ze wzoru 9. Decyduje ona czy na wyjściu otrzymana wartość to 1 czy 0.

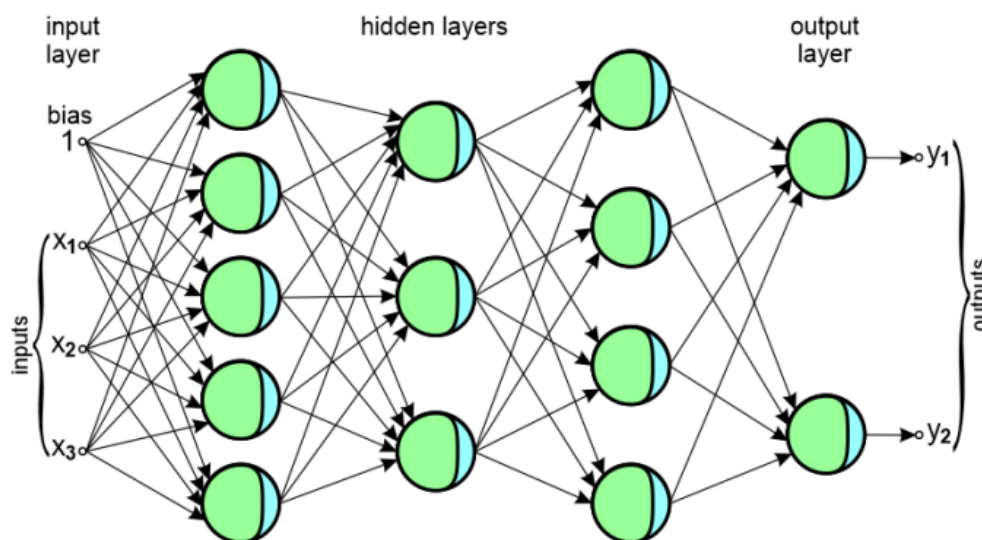
Z czasem zaczęły powstawać kolejne koncepcje budowy sieci neuronowych, między innymi perceptron wielowarstwowy (sieci MLP). Charakteryzują się one trzema typami warstw:

- Warstwa wejściowa – odpowiada za przyjmowanie neuronów i zawsze jest jedna w całej

sieci.

- Warstwy ukryte – może być jedna lub więcej i odpowiadają za tworzenie połączeń między neuronami i tworzenie kombinacji, których nie dałoby się wykonać za pomocą tylko jednej warstwy.
- Warstwa wyjściowa – jest ostatnią z warstw i zwraca wynik jaki został wytrenowany w trakcie obliczeń wykonywanych przez poprzednie warstwy. Może ona zwracać liczby, a także cechy.

Schemat budowy sieci neuronowej z trzema warstwami ukrytymi jest przedstawiony na rysunku 12. Na zdjęciu widać trzy wejścia opisane jako: x_1, x_2, x_3 oraz wcześniej opisywany bias (oznaczony 1). Następnie są trzy warstwy ukryte o liczbie neuronów odpowiednio: 5, 3 i 4. Strzałkami są zaznaczone, które wejścia są dodawane do neuronów. Następnie są one mnożone przez wagę przypisaną neuronowi sumowane z pozostałymi wejściami oraz poddawane pod funkcję aktywacji. Taki neuron na wyjściu staje się wejściem do kolejnej warstwy ukrytej. Każde wejście oraz każdy neuron z warstw ukrytych ma połączenie z każdym możliwym neuronem z kolejnej warstwy ukrytej. Jest to podejście zwane pełnym połączeniem (z ang. fully connected), gdzie wykorzystuje się wszystkie możliwe kombinacje połączeń między neuronami. Spotykane są też inne koncepcje, jednak w tej pracy wykorzystano podejście z pełnymi połączeniami.



Rysunek 12: Model sieci MLP z trzema wejściami: x_1, x_2, x_3 i jednym biasem, trzema warstwami ukrytymi oraz dwoma wyjściami y_1, y_2 . Rysunek zaczerpnięty z [31]

Sieci MLP są tylko jednym z wielu przykładów sieci neuronowych. Innymi, które w dzisiejszych czasach są rozpowszechnione są np. sieci konwolucyjne, rekurencyjne, grafowe. Z uwagi na to, że sieci neuronowe nie były dotychczas testowane do pomiarów łamania parzystości CP , to w niniejszej pracy użyto podstawowych sieci MLP i opis ograniczono do perceptronu wielowarstwowego.

Kolejnym pojęciem bez którego nie sposób mówić o sieciach neuronowych są funkcje aktywacji. Przyjmują one jako argument wartość Y ze wzoru 9. Występują we wszystkich warstwach i na początkowym etapie decydują, która wartość zostanie przekazana dalej, a w ostatniej warstwie decydują co będzie na wyjściu modelu. Nie są one konieczne, ponieważ można zostawić wartości zwracane po pomnożeniu wag z neuronami, jednak nie jest to najefektywniejszy sposób. Wywołanie funkcji aktywacji jest momentem, gdzie sieć przestaje mieć wyłącznie zależności liniowe, a to może zwiększyć jej efektywność. Często stosowanymi funkcjami aktywacyjnymi są: funkcje trygonometryczne (sinus, tangens), sigmoida, funkcja ReLU (przy argumentach większych od zera zachowuje się jak funkcja liniowa, natomiast jeśli argumenty są poniżej zera zwraca zawsze 0), a także funkcje uni- lub bipolarne. Parametrami, które również odgrywają ważną rolę w algorytmach sieci neuronowych to:

- funkcja kosztu, która jest miarą jak wytrenowane wartości modelu różnią się od rzeczywistych;
- optymalizator, który decyduje jak zmienić wartość wagi neuronu w poszczególnych iteracjach; często wykorzystywane do tego są metody gradientowe;
- liczba epok, która określa ile razy dane zostaną pokazane modelowi, czyli ile razy będą wykonywane iteracje i odpowiednio modyfikowane wagi, aby uzyskać jak najefektywniejsze wyniki; dobór tej liczby jest o tyle znaczący, że zbyt mała liczba epok może prowadzić, że efektywność modelu będzie niska, a zbyt duża z kolei może prowadzić do zjawiska nazywanego przetrenowaniem, czyli nadmiernym dopasowaniem się modelu do danych uczących.

Podsumowując, w niniejszej pracy do poszukiwania różnic między cząstkami i antycząstkami użyto trzech metod: metody S_{CP} (standardowej metody używanej w eksperymencie LHCB), algorytmu Imana-Conovera (w celu poprawy czułości metody S_{CP}) i sieci neuronowych. Zarówno algorytm Imana-Conovera, jak i sieci neuronowe nie były wcześniej używane do poszukiwań łamania parzystości CP .

4 Pomiary różnic między cząstkami i antycząstkami

W tym rozdziale przedstawiono wyniki poszukiwań różnic między cząstkami i antycząstkami otrzymane przy użyciu każdej z metod opisanej w poprzednim rozdziale. Analizowane dane zostały wcześniej wygenerowane i ich opis jest zawarty w rozdziale 2.

4.1 Metoda S_{CP}

Przy użyciu metody S_{CP} porównano dwuwymiarowe rozkłady Dalitza cząstek i antycząstek w celu znalezienia różnic między nimi. Liczbę przedziałów na jaką podzielono rozkłady Dalitza ustalono metodą Sturgesa [32]. Metoda ta pozwala na wyliczenie liczby przedziałów na podstawie rozmiaru próbki danych, korzystając z formuły:

$$n = \log_2 N + 1, \quad (10)$$

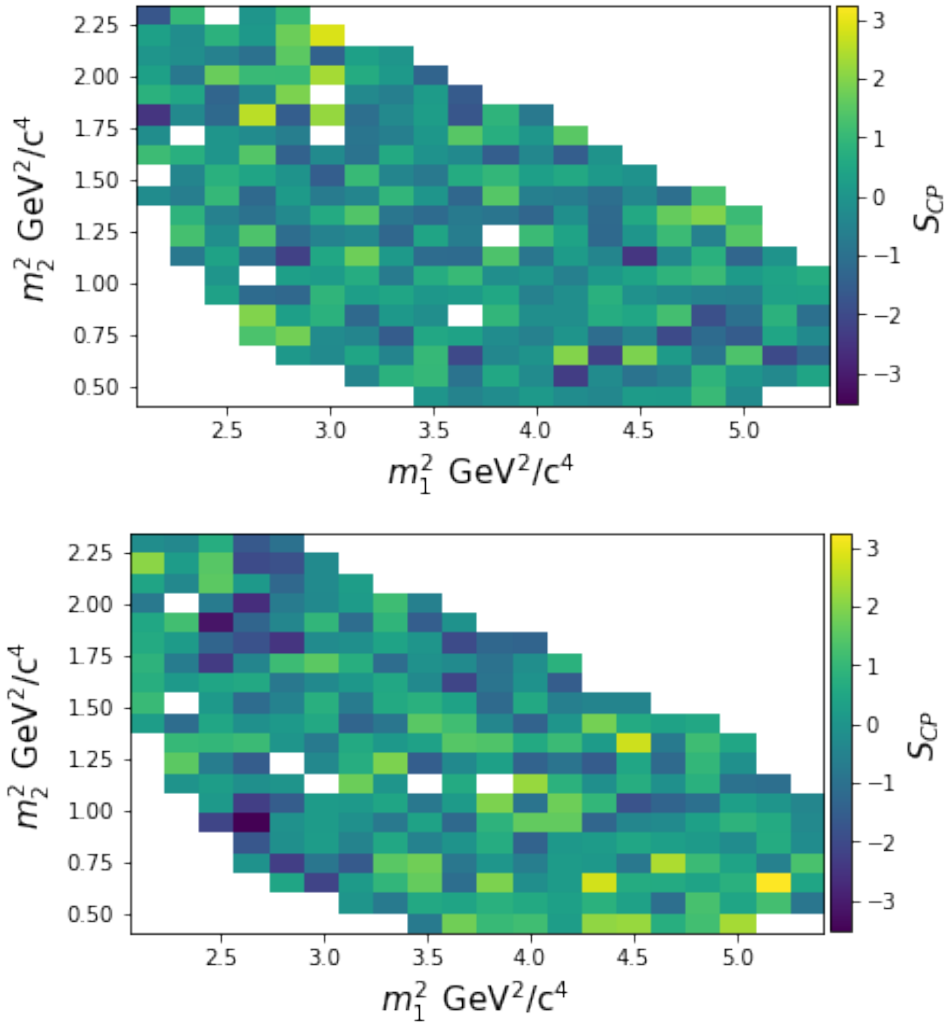
gdzie: N jest statystyką zbioru (liczbą rozpadów), a n jest liczbą przedziałów. Metoda Sturgesa zabezpiecza przed nadmiernym lub niedostatecznym agregowaniem danych w histogramie. Zbyt wiele przedziałów z niedostateczną statystyką może spowodować, że metoda będzie czuła na fluktuacje statystyczne, które mogą zostać błędnie zinterpretowane jako nowy efekt (w rozpatrywanej sytuacji jako różnice między cząstkami i antycząstkami wynikające z łamania parzystości CP). Z drugiej strony, zbyt mała liczba przedziałów może prowadzić do utraty ważnych szczegółów, które zawierają dane.

W sytuacji 100 tysięcy cząstek na histogramie jednowymiarowym podstawienie do wzoru 10 daje wynik $n \approx 20$. Dlatego osie rozkładów Dalitza cząstek i antycząstek zostały podzielone na 20 przedziałów każda z nich.

4.1.1 Wyniki otrzymane na jednej próbce

Korzystając ze wzoru 5 obliczono wartości S_{CP}^i w każdym przedziale rozkładu Dalitza. Na rysunku 13 zostały przedstawione mierzone wartości S_{CP}^i uzyskane na danych z jednej próbki bez i z asymetrią (łamaniem parzystości CP). W przypadku próbki bez asymetrii otrzymano 261 przedziałów, w których zarówno liczba cząstek, jak i antycząstek przekroczyła 10 zliczeń. Natomiast w przypadku próbki z asymetrią otrzymano 265 analogicznych przedziałów. Można zauważyć, że w przypadku próbki z asymetrią pojawia się więcej przedziałów z wyższymi bezwzględnymi wartościami S_{CP}^i , bliższymi trzem standardowym odchyleniom (zwłaszcza w le-

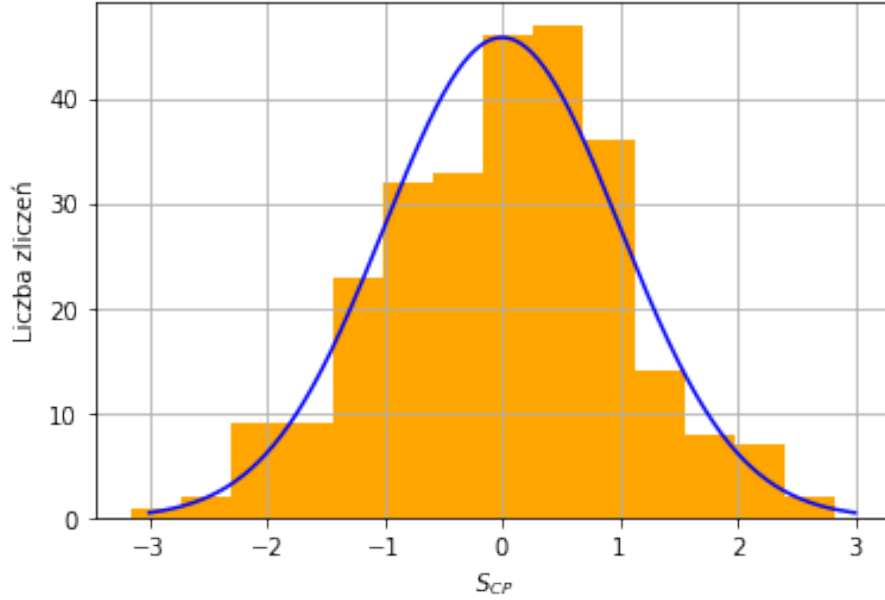
wym górnym i prawym dolnym narożniku). Sugeruje to występowanie w tych obszarach różnic między rozkładami cząstek i antycząstek.



Rysunek 13: Wartości S_{CP} w przedziałach rozkładu Dalitza uzyskane na danych z jednej próbki bez (górny rysunek) i z asymetrią (dolny rysunek). Kolor biały oznacza przedziały, gdzie nie była obliczana wartość S_{CP} , czyli te gdzie liczba cząstek i/lub antycząstek była mniejsza lub równa 10.

Rozkład mierzonych wartości S_{CP}^i w próbce bez asymetrii został przedstawiony na rysunku 14. Wartości te są od -3 do 3 standardowych odchylen z medianą w okolicy zera, a kształt ich rozkładu przypomina dzwon, co jest charakterystyczne w przypadku rozkładu normalnego. W sytuacji braku łamania parzystości CP rozkład wartości S_{CP}^i powinien być zgodny z rozkładem normalnym, o wartości oczekiwanej $\mu = 0$ i odchyleniu standardowym $\sigma = 1$. Dlatego do rozkładu wartości S_{CP}^i został dopasowany rozkład Gaussa, a mierzone μ i σ zostały podane

w tabeli 3. W granicach niepewności statystycznych mierzone μ i σ są zgodne z rozkładem normalnym. Ponadto mierzone prawdopodobieństwo zgodności rozkładów cząstek i antycząstek testem χ^2 wynosi 0,47. Pokazuje to, że metoda S_{CP} nie generuje fałszywego sygnału łamania parzystości CP , gdy nie jest ono oczekiwane.



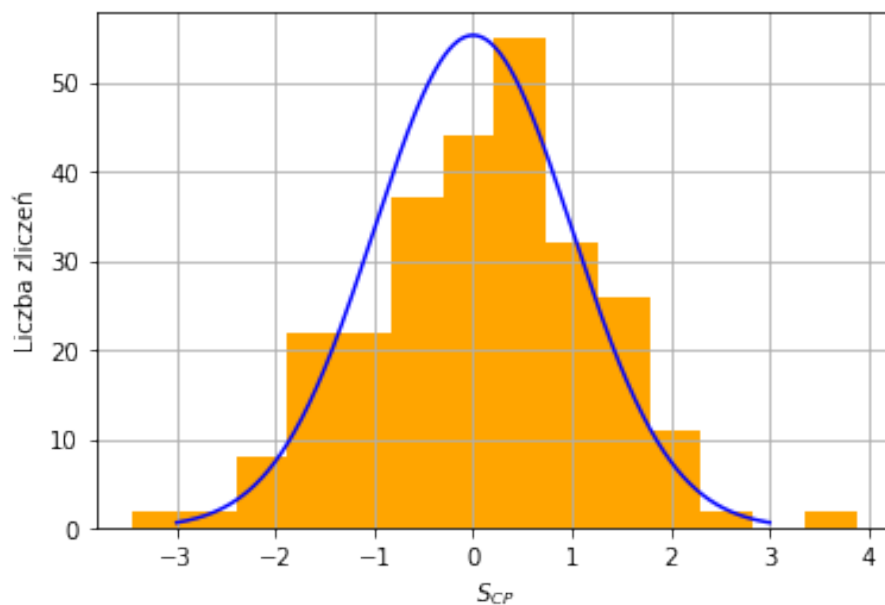
Rysunek 14: Rozkład wartości S_{CP} mierzonych w jednej próbce danych z brakiem łamania parzystości CP . Do danych (histogram pomarańczowy) dopasowany rozkład Gaussa (linia niebieska).

Tabela 3: Mierzone μ i σ wraz z niepewnościami $\Delta\mu$ i $\Delta\sigma$ z dopasowania funkcji Gaussa do danych z rysunku 14. Sytuacja z brakiem łamania parzystości CP .

μ	$\Delta\mu$	σ	$\Delta\sigma$
0,00	0,06	0,96	0,06

Analogiczne kroki wykonano na próbce danych z asymetrią. Mierzone w tym przypadku wartości S_{CP}^i w każdym przedziale rozkładu Dalitza zostały przedstawione na rysunku 15. Wartości minimalna i maksymalna S_{CP}^i są nieznacznie większe, co do wartości bezwzględnej, od tych, które zostały zmierzone w próbce bez asymetrii (rysunek 14) i przekraczają trzy standardowe odchylenia. Ponadto, w tym przypadku, rozkład S_{CP}^i nie jest już symetryczny i jego

kształt różni się od rozkładu z rysunku 14, czyli różni się od rozkładu normalnego. Wartości μ i σ otrzymane z dopasowania funkcji Gaussa do tych danych zostały podane w tabeli 4. Wartości te są zgodne w granicach niepewności statystycznych z rozkładem normalnym sugerując zgodność statystyczną rozkładów cząstek i antycząstek. Mierzone prawdopodobieństwo statystycznej zgodności tych rozkładów testem χ^2 wynosi 0,0048, co odpowiada 2,6 standardowym odchyleniom. Mimo zauważalnych różnic między rozkładami z i bez asymetrii, metodą S_{CP} nie można jednoznacznie potwierdzić istnienia wielkości włożonej asymetrii między rozkładami cząstek i antycząstek. W takiej sytuacji należałoby zwiększyć statystykę zbieranych danych. Jednakże nie zawsze jest to proste. Dlatego przy ograniczonej statystyce danych warto jest poszukiwać metod, które być może będą czulsze na badane efekty. W tej pracy przedstawione jest nowe podejście do pomiarów łamania parzystości CP z wykorzystaniem algorytmu Imana-Conovera i sieci neuronowych.



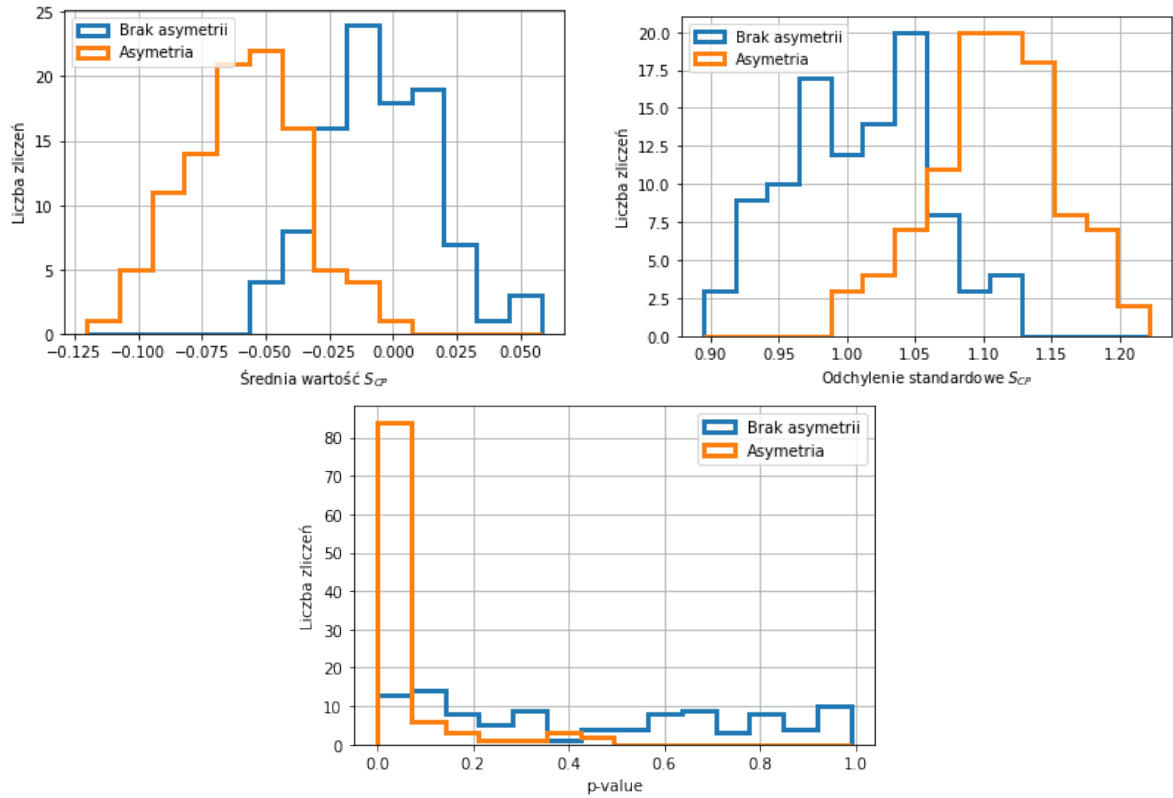
Rysunek 15: Rozkład wartości S_{CP} mierzonych w jednej próbce danych z asymetrią. Na rozkład naniesiona krzywa Gaussa o μ i σ otrzymanymi z dopasowania tej funkcji do danych bez asymetrii.

Tabela 4: Mierzone μ i σ wraz z niepewnościami $\Delta\mu$ i $\Delta\sigma$ z dopasowania funkcji Gaussa do danych z rysunku 15. Sytuacja z asymetrią, czyli niezerowym łamaniem parzystości CP wynoszącym 5% w amplitudzie stanu rezonansowego K^* .

μ	$\Delta\mu$	σ	$\Delta\sigma$
-0,04	0,07	1,09	0,07

4.1.2 Wyniki otrzymane na 100 próbkach

Z uwagi na to, że na jednej próbce danych nie można oszacować czułości metody, to w tym celu powtórzono powyższe badania na 100 próbkach, raz z brakiem łamania parzystości CP w każdej próbce, a drugi raz z łamaniem parzystości CP w każdej próbce jako 5% różnicy w amplitudzie stanu rezonansowego K^* (tzw. próbki z asymetrią). Na rysunku 16 przedstawiono średnie wartości S_{CP} i odpowiadające im odchylenie standardowe oraz prawdopodobieństwo zgodności rozkładów cząstek i antycząstek uzyskane metodą S_{CP} w każdej ze 100 próbek z i bez asymetrii. Średnie wartości wraz z niepewnościami średniej wartości S_{CP} i jego odchylenia standardowego zostały podane w tabeli 5. W sytuacji braku asymetrii w próbkach, rozkład średnich wartości S_{CP} jest symetryczny wokół zera (w granicach niepewności statystycznej) w zakresie od $-0,05$ do $0,05$. Rozkład odchylenia standardowego tej średniej również jest symetryczny wokół jedynki (w granicach niepewności statystycznej) i przyjmuje wartości od $0,9$ do $1,1$. Natomiast w sytuacji próbek z asymetrią analogiczne rozkłady są przesunięte. Rozkład średnich S_{CP} jest przesunięty w kierunku ujemnych wartości ze średnią $-0,06 \pm 0,07$. Rozkład odchylenia standardowego tej średniej jest przesunięty w kierunku wyższych wartości niż jeden z medianą wynoszącą $1,11 \pm 0,07$. Pokazuje to wyraźne różnice między próbkami z i bez asymetrii. Różnice te są także obserwowane w rozkładach prawdopodobieństwa (p-value). W sytuacji braku asymetrii rozkład p-value jest płaski. Natomiast w sytuacji istnienia niezerowej asymetrii ten rozkład nie jest już płaski i wartości koncentrują się w okolicach zera. Wskazuje to na statystyczne różnice między rozkładami cząstek i antycząstek. Ten test pokazuje, że metoda S_{CP} nie generuje fałszywych sygnałów łamania parzystości CP , gdy ono nie występuje i może potwierdzić istnienie tego efektu, gdy jest ono na poziomie 5% różnicy w amplitudzie rezonansu K^* .



Rysunek 16: Rozkłady średnich wartości S_{CP} (lewy górny rysunek), odchylenia standardowego średniej (prawy górny rysunek) i prawdopodobieństwa zgodności rozkładów cząstek i antycząstek – p-value (dolny rysunek) mierzone metodą S_{CP} na 100 próbkach danych z brakiem asymetrii (linia niebieska) i z asymetrią (linia pomarańczowa).

Tabela 5: Średnie wraz z niepewnościami średnich wartości S_{CP} (μ , $\Delta\mu$) i ich odchylenia standardowego (σ , $\Delta\sigma$) uzyskane na 100 próbkach danych z łamaniem parzystości CP (próbki z asymetrią) i jej brakiem (próbki bez asymetrii).

Łamanie CP	μ	$\Delta\mu$	σ	$\Delta\sigma$
5%	-0,06	0,07	1,11	0,07
Brak	-0,01	0,06	1,01	0,06

4.1.3 Czułość metody

Sprawdzono ile razy na 100 prób mierzone jest prawdopodobieństwo zgodności rozkładów cząstek i antycząstek na poziomie powyżej 3 i 5 standardowych odchylen. Wyniki zostały przedstawione w tabeli 6. Przy statystyce 200 000 przypadków (łącznie cząstek i antycząstek)

na 100 prób metoda S_{CP} w żadnej próbie nie potwierdziła istnienia różnic między rozkładami cząstek i antycząstek (asymetrii – łamania parzystości CP wynoszącego 5% w amplitudach stanu rezonansowego K^*), czyli prawdopodobieństwa ich zgodności przekraczającego pięć standardowych odchyłeń. Oznacza to, że skuteczność metody przy badanej statystyce i przyjętej wielkości różnicy między porównywanymi zbiorami wynosi 0%. Natomiast ewidencja niezgodności rozkładów cząstek i antycząstek została zmierzona w 31% prób. Dlatego w celu zwiększenia skuteczności metody S_{CP} , czyli jej poprawy w odnajdywaniu asymetrii między porównywanymi próbkami, w niniejszej pracy użyto algorytmu Imana-Conovera. Wyniki tych badań zostały przedstawione w następnym rozdziale.

Tabela 6: Liczba próbek, w których mierzone jest prawdopodobieństwo zgodności rozkładów cząstek i antycząstek (p-value) powyżej 3 i 5 standardowych odchyłeń (σ).

Łamanie CP	p-value powyżej 3σ	p-value powyżej 5σ
Brak	0	0
5%	31	0

4.2 Algorytm Imana-Conovera

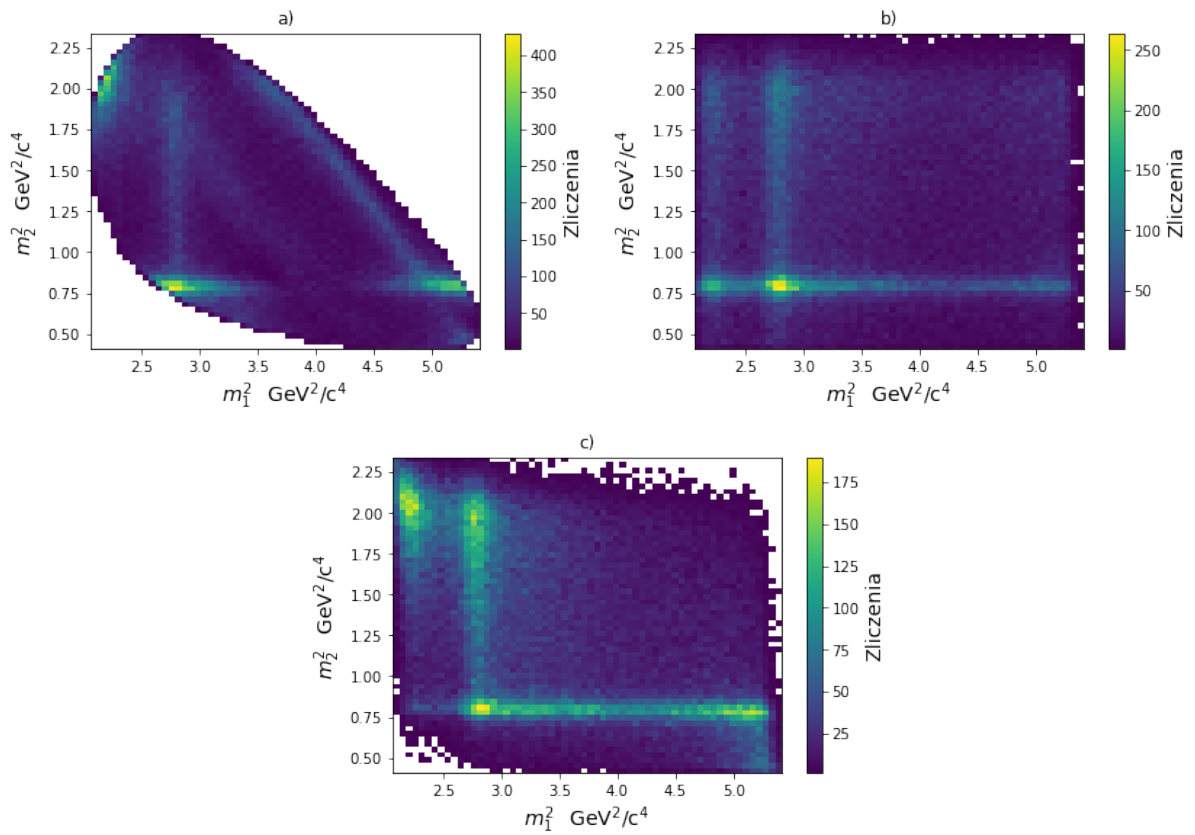
Algorytm Imana-Conovera został użyty w celu sprawdzenia, czy jest on w stanie zwiększyć czułość metody S_{CP} , której wyniki zostały przedstawione w poprzednim rozdziale. W tym celu, przed zastosowaniem metody S_{CP} , dane zostały poddane transformacji przez algorytm Imana-Conovera.

4.2.1 Transformacja danych

Na danych zastosowano dekorelację, znormalizowano je i poddano transformacji Imana-Conovera. Wizualny efekt poszczególnych przekształceń rozkładu Dalitza został przedstawiony na rysunku 17. W wyniku dekorelacji zmiennych, a następnie transformacji Imana-Conovera więcej przedziałów na rozkładzie Dalitza jest niepustych (z minimalną liczbą zliczeń cząstek i antycząstek większą od 10 każda) w związku z innym rozłożeniem się statystyki (szczegółowa dyskusja jak zmieniła się liczba przedziałów podana jest w rozdziale 4.2.2). Może to mieć znaczenie w metodzie S_{CP} , gdyż potencjalnie może to pomóc w zwiększeniu czułości testu χ^2 . Daje to nadzieję na zwiększenie liczby przedziałów, w których pojawią się różnice między cząstkami

i antycząstkami. Maksymalna liczba zliczeń w przedziałach nie przekracza teraz 1 200. Przed transformacją maksymalna liczba zliczeń w danym przedziale była wyższa, ale nie przekraczała 1 600.

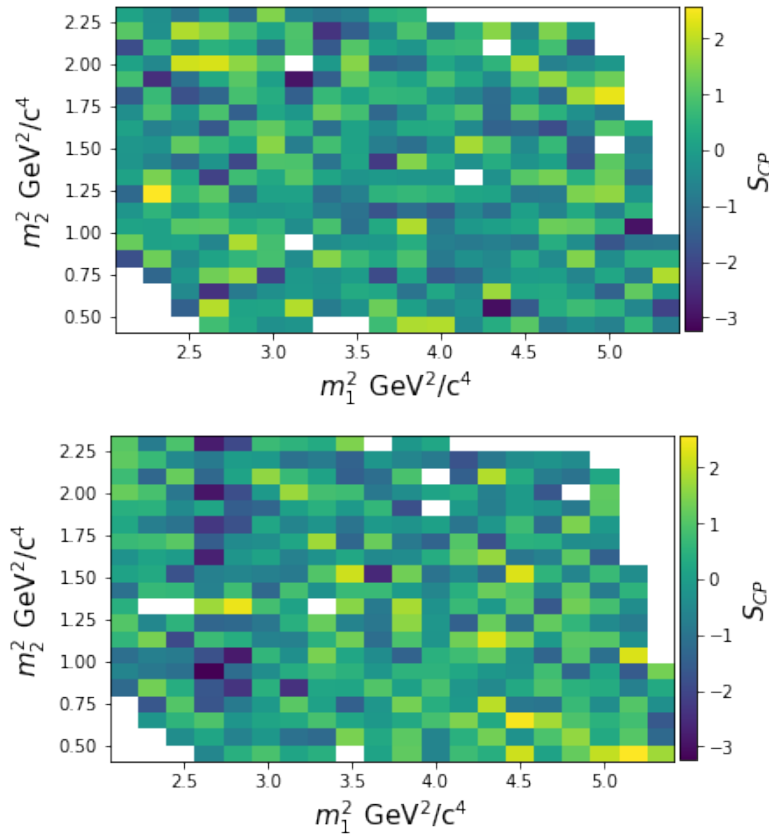
Algorytm Imana-Conovera opiera się na korelacjach między zmiennymi. Jako że na wejściu były dwie zmienne to korelacja mogła występować tylko między nimi i wynosiła $-0,57$. Po transformacji ta korelacja wynosi $-0,52$. W przypadku korelowania algorytmem Imana-Conovera drobne rozbieżności w wyniku mogą występować ze względu na specyfikę metody (zostało to opisane w rozdziale 3.2).



Rysunek 17: Proces zmian na rozkładzie Dalitza w wyniku działania algorytmu Imana-Conovera na przykładzie jednej próbki z brakiem łamania parzystości CP (bez asymetrii): oryginalny rozkład Dalitza – dane wejściowe (a), dekorelacja (b), dane po transformacji Imana-Conovera (c). W celu wizualizacji zachodzących procesów obie osie rozkładu Dalitza zostały podzielone na 60 przedziałów każda.

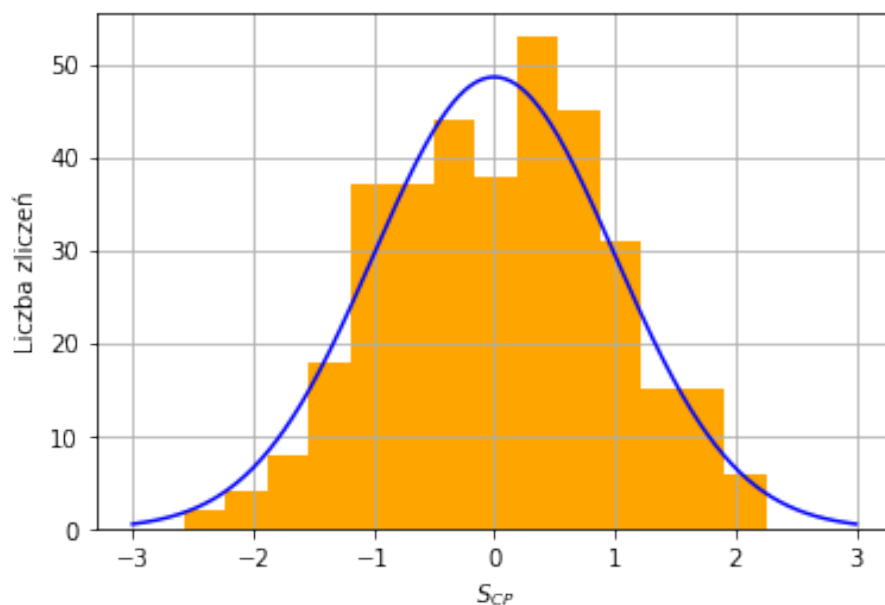
4.2.2 Wpływ na czułość metody S_{CP}

Na danych zmienionych w wyniku działania algorytmu Imana-Conovera zastosowano metodę S_{CP} w identyczny sposób jak opisany w rozdziale 4.1. Obie osie rozkładu Dalitza zostały podzielone na identyczną liczbę przedziałów, czyli po 20. Mierzone teraz wartości S_{CP}^i w poszczególnych przedziałach (w tej samej próbce danych bez asymetrii i z asymetrią, jak poprzednio) zostały przedstawione na rysunku 18. Teraz liczba przedziałów z minimalną liczbą zliczeń jest większa i wynosi 361 (poprzednio 261) w próbce bez asymetrii oraz 350 (poprzednio 265) w próbce z asymetrią. Oznacza to, że liczba przedziałów zwiększyła się o około 35%. Maksymalna liczba zliczeń w przedziałach zmniejszyła się do 1 449 (poprzednio 1 804) w próbce bez asymetrii i do 1 434 (poprzednio 1 893) w próbce z asymetrią. Ta względnie duża statystyka danych w przedziałach nadal powinna zapewnić, że metoda S_{CP} nie będzie czuła na fluktuacje statystyczne. Można też zauważyć, że teraz w próbce z asymetrią w obszarze koło $m_1^2 \approx 2,7 \text{ GeV}^2/c^4$ jest więcej wartości S_{CP}^i bliskich -3 .

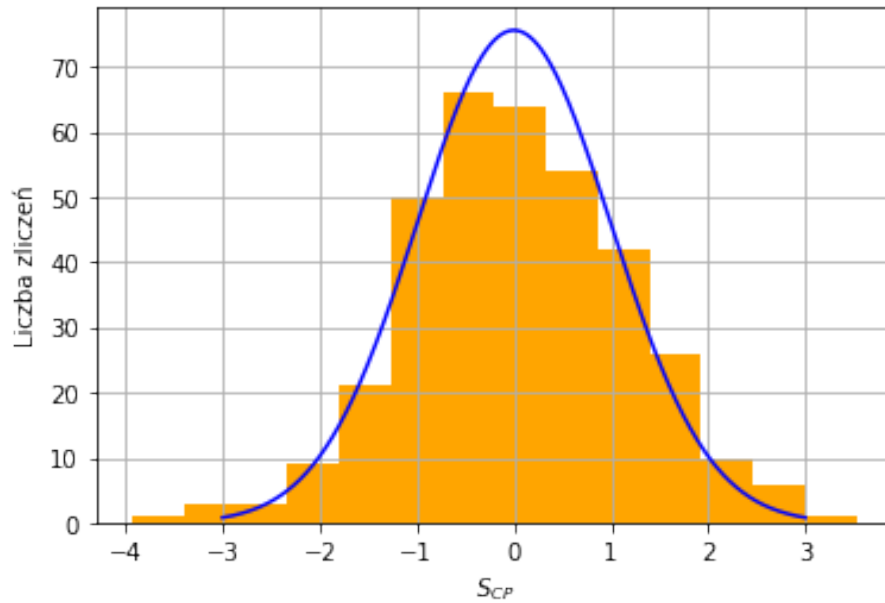


Rysunek 18: Mierzone wartości S_{CP} w przedziałach rozkładu Dalitza w jednej próbce danych bez asymetrii (górny rysunek) i z asymetrią (dolny rysunek) po zastosowaniu algorytmu Imana-Conovera.

Na rysunku 19 został przedstawiony rozkład wartości S_{CP}^i mierzony w próbce bez asymetrii, a na rysunku 20 uzyskany z próbki z asymetrią. Sprawdzono czy $\mu=0$ oraz $\sigma=1$, a wyniki w tabeli 7 pokazują, że tak jest, choć w przypadku odchylenia standardowego jest ono zwiększone dla próbki z asymetrią o około 6%. W przypadku próbki z asymetrią pojawia się nieznacznie więcej przedziałów z wartościami od -3 do -2 niż bez użycia algorytmu Imana-Conovera przedstawionego na rysunku 15 (nieznacznie więcej zliczeń powyżej rozkładu normalnego). Jest to pozytywny efekt, który daje nadzieję na zwiększenie czułości metody S_{CP} . W celu oceny tej czułości powtórzono pomiary na 100 próbkach.



Rysunek 19: Rozkład mierzonych wartości S_{CP} w jednej próbce bez asymetrii po transformacji Imana-Conovera. Do danych (histogram pomarańczowy) dopasowano rozkład Gaussa (linia niebieska).



Rysunek 20: Rozkład mierzonych wartości S_{CP} w jednej próbce z asymetrią po transformacji Imana-Conovera. Dodatkowo naniesiono rozkład Gaussa, który był otrzymany z dopasowania do danych bez asymetrii.

Tabela 7: Mierzone μ i σ wraz z niepewnościami $\Delta\mu$ i $\Delta\sigma$ z dopasowania funkcji Gaussa do danych z rysunków 19 oraz 20. Porównywana jest sytuacja z asymetrią, a także ta w której ona nie występuje.

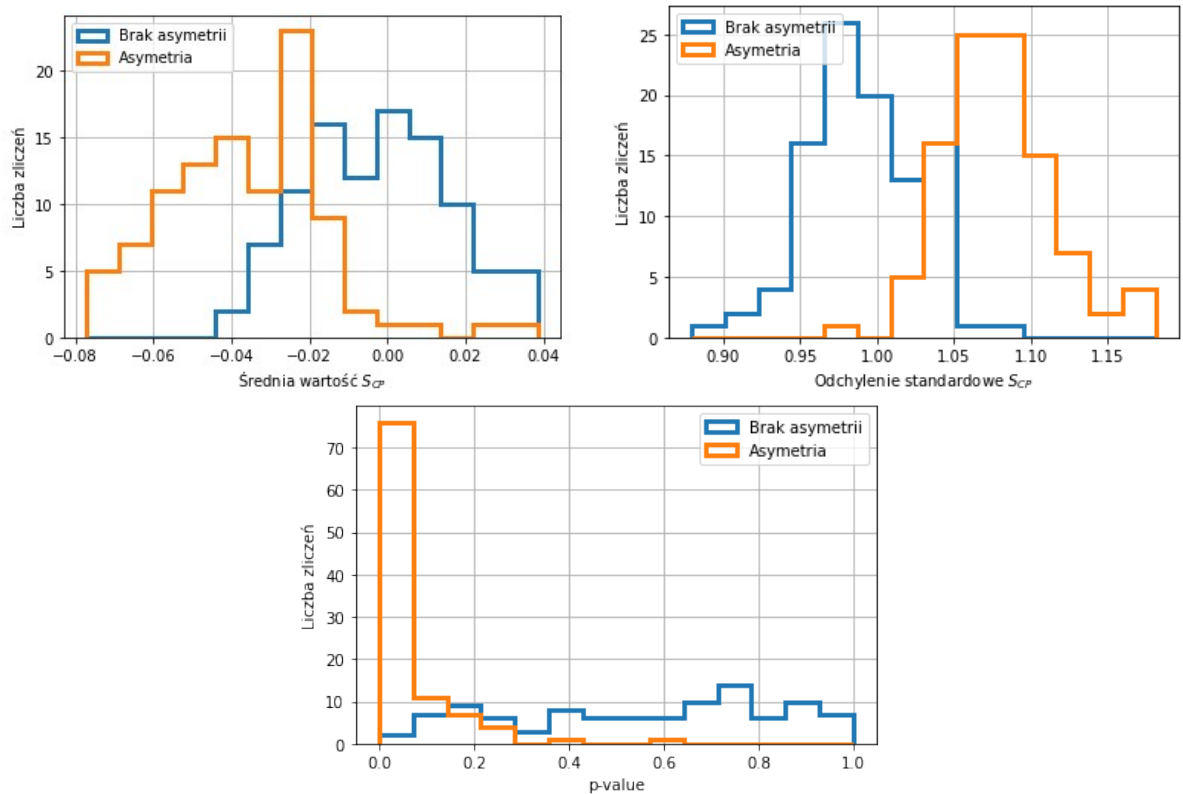
Łamanie CP	μ	$\Delta\mu$	σ	$\Delta\sigma$
5%	-0,03	0,06	1,07	0,06
Brak	-0,01	0,06	1,01	0,06

Analogicznie jak poprzednio, na rysunku 21 przedstawiono średnie wartości S_{CP} i odpowiadające im odchylenie standardowe oraz prawdopodobieństwo zgodności rozkładów cząstek i antycząstek uzyskane metodą S_{CP} w każdej ze 100 próbek z i bez asymetrii, ale tym razem po uwzględnieniu przekształceń danych algorytmem Imana-Conovera. Natomiast średnie wartości wraz z niepewnościami średniej wartości S_{CP} i jej odchylenia standardowego zostały podane w tabeli 8. Także tutaj obserwuje się analogiczne różnice między próbkami bez i z asymetrią. W przypadku próbek bez asymetrii rozkład wartości S_{CP} jest symetryczny wo-

kół zera (w granicach niepewności statystycznej) z minimalną wartością $-0,04$ i maksymalną $0,04$. Odpowiadający mu rozkład uzyskany na danych z asymetrią jest przesunięty w kierunku ujemnych wartości ze średnią $-0,04 \pm 0,06$. Co prawda, ta wartość jest zgodna z zerem, ale rozkład jest wyraźnie asymetryczny i minimalna wartość jest niższa i wynosi $-0,08$.

Jeśli zaś chodzi o odchylenie standardowe średniej, to w przypadku próbek bez asymetrii jego średnia wartość jest zgodna z 1 (w granicach niepewności statystycznej), tak jak poprzednio. Rozkład tych wartości wydaje się być też symetryczny wokół 1 z minimalną wartością 0,9 i maksymalną 1,1. Natomiast w przypadku próbek z asymetrią rozkład jest wyraźnie przesunięty w kierunku wartości wyższych ze średnią $1,08 \pm 0,06$. Oznacza to, że rozkład jest ok. 8% przesunięty w porównaniu do rozkładu bazowego (bez asymetrii). Maksymalne wartości odchylenia standardowego też są wyższe i sięgają do 1,15.

W przypadku rozkładu prawdopodobieństwa statystycznej zgodności rozkładów cząstek i antycząstek także obserwuje się charakterystyczne różnice. Rozkład ten jest płaski w przypadku próbek bez asymetrii, co oznacza, że po działaniu algorytmem Imana-Conovera na dane metoda S_{CP} nie generuje fałszywych sygnałów różnic między cząstkami i antycząstkami. Natomiast w przypadku próbek z asymetrią rozkład wartości prawdopodobieństwa nie jest płaski, a mierzone wartości p-value są w większości bliskie zeru. Oznacza to, że hipoteza zerowa nie jest spełniona (hipoteza zerowa oznacza, że rozkłady cząstek i antycząstek są tożsame, a rozkład wartości S_{CP}^i w każdej próbce jest zgodny z rozkładem normalnym). Podsumowując, algorytm Imana-Conovera nie generuje fałszywych wyników i potwierdza zjawisko łamania parzystości CP , gdy jest ono uwzględnione w danych na założonym wcześniej poziomie.



Rysunek 21: Rozkłady średnich wartości S_{CP} (lewy górny rysunek), odchylenia standardowego średniej (prawy górny rysunek) i prawdopodobieństwa zgodności rozkładów cząstek i antycząstek – p-value (dolny rysunek) mierzone metodą S_{CP} z uwzględnieniem algorytmu Imana-Conovera na 100 próbkach danych z brakiem asymetrii (linia niebieska) i z asymetrią (linia pomarańczowa).

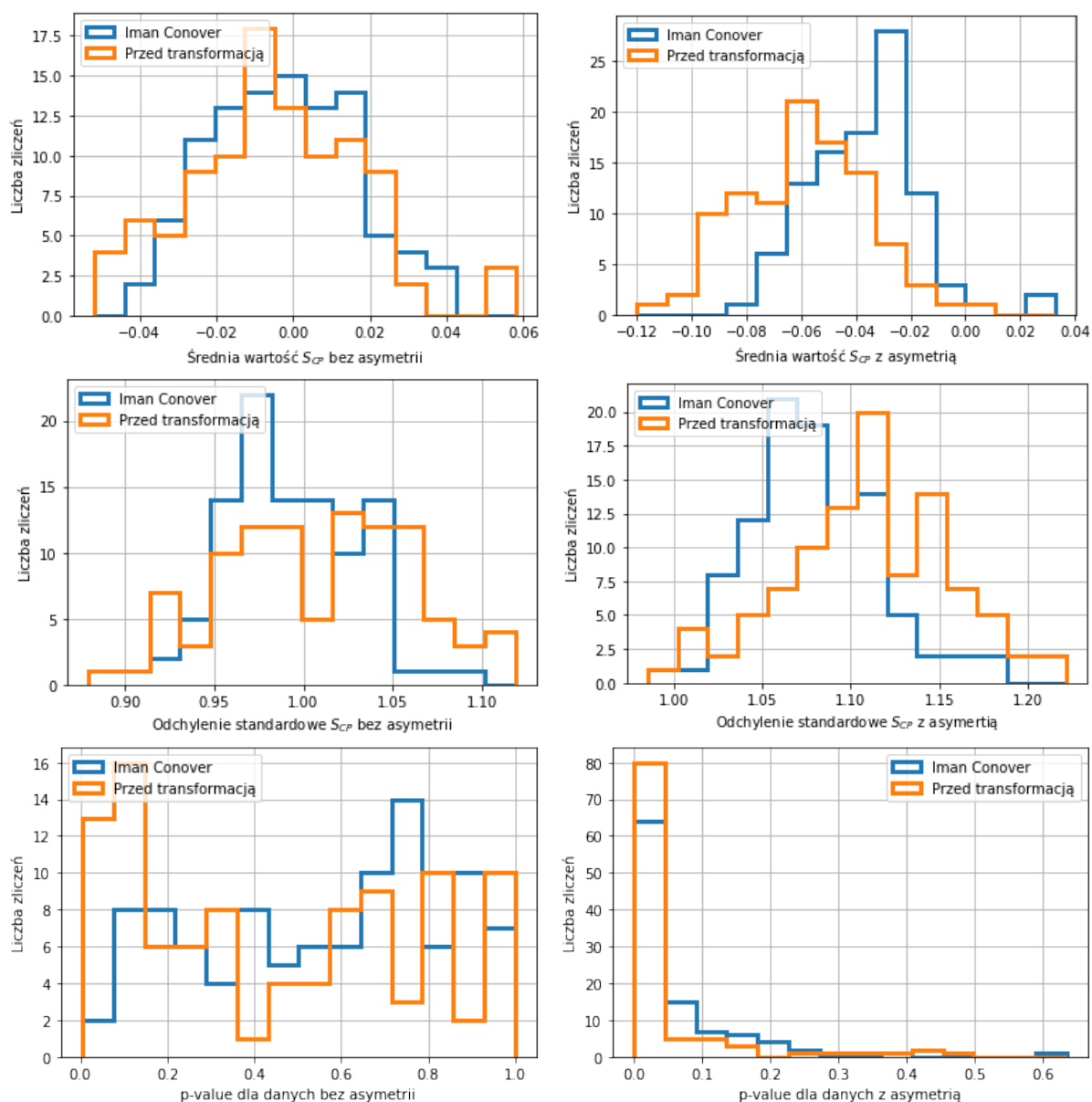
Tabela 8: Średnie wraz z niepewnościami średnich wartości S_{CP} (μ , $\Delta\mu$) i ich odchylenia standardowego (σ , $\Delta\sigma$) uzyskane na 100 próbkach danych z łamaniem parzystości CP (próbki z asymetrią) i jej brakiem (próbki bez asymetrii) po uwzględnieniu algorytmu Imana-Conovera.

Łamanie CP	μ	$\Delta\mu$	σ	$\Delta\sigma$
5%	-0,04	0,06	1,08	0,06
Brak	-0,00	0,06	1,01	0,06

W celu sprawdzenia, czy użycie algorytmu Imana-Conovera poprawia czułość metody S_{CP} porównano wyniki otrzymane tą metodą przed i po zastosowaniu algorytmu, czyli wyniki z rysunku 16 z rysunkiem 21. Rezultaty tych porównań zostały przedstawione na rysunku 22. W przypadku braku asymetrii (rysunki po lewej stronie) rozkłady danych przed i po zastosowaniu

waniu algorytmu Imana-Conovera mają podobne kształty. Co prawda rozkład p-value może wydawać się być inny w pierwszych dwóch przedziałach najmniejszych wartości. W pierwszym przedziale przed transformacją Imana-Conovera liczba zliczeń wynosi 13, a po zastosowaniu algorytmu Imana-Conovera ta liczba jest mniejsza i wynosi 2. Jednakże te liczby są ze sobą zgodne na granicy trzech standardowych odchyłeń. W kolejnym przedziale problem nie jest już tak zauważalny. Tutaj przed transformacją Imana-Conovera liczba zliczeń wynosi 15, a po transformacji 8. Te wartości są tożsame w granicach dwóch standardowych odchyłeń. Ponadto biorąc pod uwagę wszystkie rozkłady średnia wartość zliczeń w przedziale wydaje się być około 7. Rozkład p-value został podzielony na 14 przedziałów. Równomierny rozdział statystyki 100 próbek na przyjętą liczbę przedziałów daje średnią liczbę prób w każdym przedziale właśnie 7. Ta wartość jest na granicy statystycznego minimum. Najlepiej byłoby, gdyby rozkład p-value został podzielony na mniej niż 10 przedziałów, co dawałoby minimum 10 prób w każdym przedziale. Jednakże zdecydowano się przyjąć nieznacznie większą liczbę przedziałów, kosztem balansowania na granicy minimum statystycznego, w celu lepszego zobrazowania najmniejszych wartości prawdopodobieństwa, które świadczą o występujących różnicach między cząstkami i antycząstkami. Liczba przedziałów w rozkładzie p-value jest identyczna w przypadku próbek bez asymetrii, jak i z asymetrią. We wszystkich przedziałach (zarówno pierwszym i drugim) liczby zliczeń zgadzają się ze średnią 7 w przypadku próbek bez asymetrii. Ten test raz jeszcze pokazuje, że transformacja Imana-Conovera nie wprowadza dodatkowych asymetrii między porównywanymi zbiorami danych (w tym przypadku rozkładów cząstek i antycząstek).

Inaczej sytuacja przedstawia się w przypadku próbek z asymetrią (rysunki z prawej strony). Tutaj obserwowane są różnice między wynikami uzyskani bez i z użyciem algorytmu Imana-Conovera. Bez transformacji Imana-Conovera średnia wartość S_{CP} wynosi $-0,06 \pm 0,07$ (tabela 5) i główna wartość jest bardziej ujemna niż po zastosowaniu przedmiotowej transformacji, gdzie wynosi ona $-0,04 \pm 0,06$ (tabela 8). Podobnie rzecz ma się w przypadku odchylenia standardowego. Centralna średnia wartość tego odchylenia jest wyższa ($1,11 \pm 0,07$) przed zastosowaniem algorytmu Imana-Conovera niż po, gdzie wynosi $1,08 \pm 0,06$. Jednakże w granicach niepewności statystycznej wartości te są ze sobą zgodne, jak również średnia wartość S_{CP} zgodna z zerem, a odchylenie standardowe zgodne z jeden (takie wartości oznaczają brak różnic między rozkładami cząstek i antycząstek). Można zatem wnioskować, że zastosowanie algorytmu Imana-Conovera może zmniejszać czułość metody S_{CP} , a nie jak się potencjalnie wydawało poprawiać.



Rysunek 22: Porównanie średnich wartości S_{CP} (najwyższy rząd), odchylenia standardowego średniej (rząd środkowy) i prawdopodobieństwa zgodności rozkładów cząstek i antycząstek – p-value (dolny rząd) mierzone metodą S_{CP} bez użycia algorytmu Imana-Conovera (linia pomarańczowa) i po zastosowaniu algorytmu Imana-Conovera (linia niebieska) na 100 próbkach danych z brakiem asymetrii (lewa kolumna) i z asymetrią (prawa kolumna).

Sprawdzono również w ilu próbach na 100 mierzone jest prawdopodobieństwo zgodności rozkładów cząstek i antycząstek na poziomie powyżej trzech i pięciu standardowych odchyłeń. Wyniki zamieszczono w tabeli 9. Po zastosowaniu algorytmu Imana-Conovera, prawdopodobieństwo zgodności porównywanych rozkładów przekraczające trzy standardowe odchylenia mierzone jest w 13 próbach na 100. Bez użycia tego algorytmu liczba pozytywnych prób jest wyższa i wynosi 31. Co prawda wartości te są ze sobą zgodne na granicy trzech standardowych odchyłeń ($13,0 \pm 3,6$, $31,0 \pm 16,8$), jednakże mogą także sugerować, że użycie algorytmu Imana-Conovera może zmniejszać czułość metody S_{CP} , a nie ją zwiększać.

Jeśli zaś chodzi o pozytywną liczbę prób przekraczającą pomiar różnic między próbkami na poziomie pięciu standardowych odchyłeń, to wynosi ona zero zarówno bez, jak i z użyciem algorytmu Imana-Conovera. To oznacza, że algorytm ten nie zmienia czułości metody S_{CP} na poziomie pięciu standardowych odchyłeń, gdzie pozytywne wyniki byłyby interpretowane jako sygnał badanego efektu (w tym przypadku potwierdzenia różnic między cząstkami i antycząstkami wynikającymi z łamania parzystości CP).

Podsumowując, z uwagi na niepewności pomiarowe uzyskane wyniki nie pozwalają jednoznacznie stwierdzić, czy dodatkowe zastosowanie algorytmu Imana-Conovera ma pozytywny, czy też negatywny wpływ na metodę S_{CP} . Wyniki te są zarówno zgodne ze stwierdzeniem, że użycie tego algorytmu ani nie psuje, ani nie poprawia czułości badanej metody. Wyniki mogą też sugerować, że zastosowanie algorytmu zmniejsza czułość metody, ale tylko na poziomie granicy statystycznej trzech standardowych odchyłeń. W obszarze powyżej pięciu standardowych odchyłeń użycie algorytmu nie zmienia czułości.

Mimo nieco pesymistycznych wniosków użycie transformacji Imana-Conovera dostarczyło cennych informacji. Fakt, że mimo wymieszania poszczególnych wartości w próbkach przy braku asymetrii nie powstały żadne sztuczne zakłócenia, a w sytuacji z asymetrią zostały one wykryte pozwala stwierdzić, że algorytm Imana-Conovera radzi sobie z wykrywaniem asymetrii lokalnych. W trakcie wykonywania analizy zauważono również ciekawą sytuację, że w przypadku gdy próbki były mniej liczne i wynosiły po 15 000 cząstek i 15 000 antycząstek w jednej próbce, to wielkość uwzględnionej asymetrii była wykrywana na wyjątkowo podobnym poziomie przed i po transformacji Imana-Conovera. Ten niespodziewany wniosek sugeruje, że warto wykonać identyczne testy, ale ze znacznie mniejszymi statystykami danych. Być może jest jakaś graniczna statystyka poniżej, której algorytm Imana-Conovera mógłby poprawić czułość testu χ^2 . Byłaby to cenna metoda przy poszukiwaniu różnic między zbiorami danych z niedostateczną

statystyką danych. W eksperymencie LHCb jest mierzonych wiele rzadkich rozpadów cząstek, w których model standardowy przewiduje niezerowe łamanie parzystości CP . Niedostateczna statystyka danych nie pozwala na zastosowanie metody S_{CP} . Gdyby udało się zwiększyć czułość tej metody w przypadku małych statystyk danych, to otworzyłoby to nowe możliwości pomiarów.

Wobec niejednoznacznego powodzenia z użycia algorytmu Imana-Conovera w tej pracy przedstawiono także inny sposób porównywania dwóch próbek danych, oparty na sieciach neuronowych.

Tabela 9: Liczba próbek, w których mierzone jest prawdopodobieństwo zgodności rozkładów cząstek i antycząstek (p-value) powyżej 3 i 5 standardowych odchyłeń (σ) po zastosowaniu algorytmu Imana-Conovera. Przekopowano także wyniki z tabeli 6 otrzymane bez użycia algorytmu Imana-Conovera.

Łamanie CP	Iman-Conover	p-value powyżej 3σ	p-value powyżej 5σ
Brak	Nie	0	0
Brak	Tak	0	0
5%	Nie	31	0
5%	Tak	13	0

4.3 Sztuczne sieci neuronowe

Wobec braku poprawy czułości metody S_{CP} przez zastosowanie algorytmu Imana-Conovera zdecydowano się na zastosowanie całkowicie nowatorskiego podejścia do poszukiwania różnic między cząstkami i antycząstkami wynikającymi z łamania parzystości CP . W tym celu podjęto się wyzwania poszukiwania sygnałów różnic między cząstkami i antycząstkami z wykorzystaniem innowacyjnych metod inteligencji obliczeniowej (powszechnie nazywanych uczeniem maszynowym). Użyto sztucznych sieci neuronowych, których opis został umieszczony w rozdziale 3.3, do zbudowania modelu uczenia maszynowego.

4.3.1 Sposób postępowania

W celu wypracowania sposobu postępowania z użyciem sieci neuronowych do rozpoznawania różnic między dwoma zbiorami danych próbowano różnych podejść, między innymi przyjęcia każdej cząstki i antycząstki jako osobnej zmiennej w postaci par obydwu kwadratów mas z rozkładu Dalitza. To rozwiązanie nie sprawdziło się, ponieważ każda z cząstek (lub antycząstek) odpowiadałaby za jeden parametr w modelu. Na podstawie jednego rozpadu (jednej pary wartości m_1^2 , m_2^2) nie ma możliwości zidentyfikowania, czy jest to cząstka czy też antycząstka. Trzeba było znaleźć metodę, która odpowiednio pogrupuje dane w taki sposób, aby różnice między cząstkami i antycząstkami były możliwe do wychwycenia przez model. Dane z każdej próbki zostały dodatkowo podzielone (pogrupowane) na mniejsze zbiory (w celu odróżnienia tych zbiorów od próbek danych, w dalszej części pracy są one nazywane paczkami danych). W każdej paczce zachowano (na wzór metody S_{CP}) dwuwymiarowy rozkłady Dalitza jako dane wejściowe do modelu. Zatem ostatecznie pojedyncza paczka danych składała się z rozkładu dwuwymiarowego (jego odpowiednikiem była macierz dwuwymiarowa, która za wartości przyjmowała liczbę zliczeń w danym przedziale rozkładu) oraz informacją czy dany przypadek był cząstką czy antycząstką. Ten sposób traktowania danych pozwala na wytrenowanie modelu w celu klasyfikacji przypadku jako cząstka lub antycząstka. Wynik efektywności tej klasyfikacji użyto do poszukiwania różnic między rozkładami cząstek i antycząstek. W sytuacji braku różnic między porównywanymi zbiorami danych efektywność klasyfikacji powinna być całkowicie losowa i wynosić 50% (czyli w przypadku próbek bez łamania parzystości CP). W sytuacji istnienia różnic między porównywanymi zbiorami danych efektywność klasyfikacji powinna być wyższa niż 50% (czyli w przypadku próbek z łamaniem parzystości CP). Związane jest to z tym, że jeżeli są różnice między porównywanymi klasami (cząstek i antycząstek) to model otrzymuje więcej szczegółów o tych klasach niż w przypadku braku różnic i tym samym powinien nauczyć się lepiej je rozpoznawać. Zatem wyższa efektywność modelu może być interpretowana jak potwierdzenie występowania asymetrii między cząstkami i antycząstkami.

Każda ze 100 próbek została podzielona na dwa zbiory: treningowy i testowy. Zbiór treningowy został użyty do wytrenowania modelu, a zbiór testowy do oszacowania efektywności jego klasyfikacji. Proporcje statystyk obu zbiorów zostały zoptymalizowane ze względu na jak najwyższą efektywność klasyfikacji przypadków z asymetrią między cząstkami i antycząstkami i są szczegółowo opisane w kolejnych rozdziałach. Dodatkowo zarówno zbiór testowy jak i treningowy zostały dalej podzielone na wspomniane wcześniej paczki danych. Ten podział

także został zoptymalizowany pod względem jak najwyżej efektywności klasyfikacji modeli i również jest opisany w kolejnych rozdziałach. Jak się okazało istotny wpływ na końcowe wyniki miał dobór liczby przedziałów na rozkładach Dalitza z danymi wejściowymi, podobnie jak w przypadku metody S_{CP} .

4.3.2 Budowa modelu

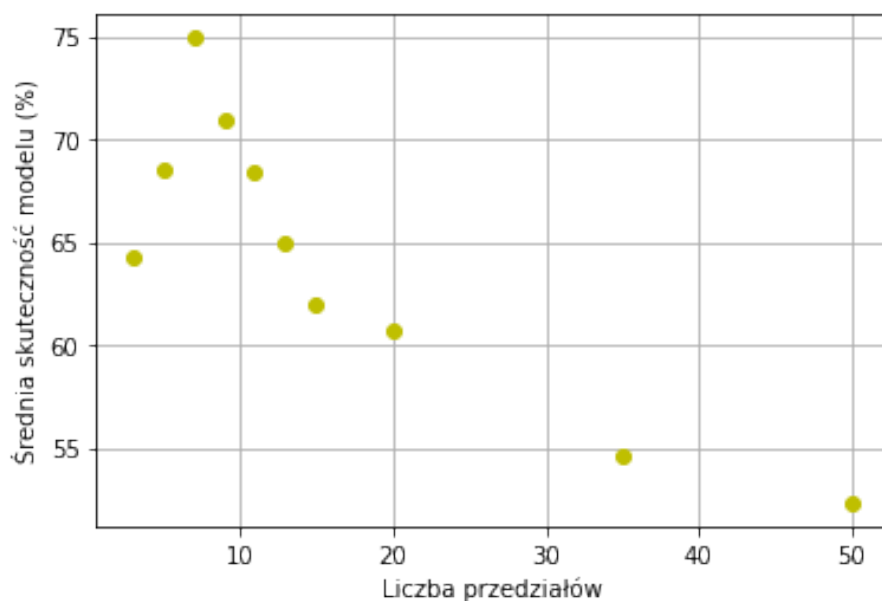
Do technicznej budowy modelu użyto język programowania Python [33], bibliotekę Tensor-Flow [21] i model sieci MLP (opisany w rozdziale 3.3). Stworzono 100 modeli z identycznymi parametrami (każda ze 100 próbek miała przyporządkowany jeden model). Niektóre parametry wyznaczano eksperymentalnie, a inne opierając się na cechach danych. Proces rozpoczęto od zamiany dwuwymiarowego rozkładu Dalitza do jednowymiarowego wektora (potocznie nazywane spłaszczenie rozkładów). Związane jest to z specyfiką zaprojektowanych sieci neuronowych [34] w używanej bibliotece. Ułatwia to przepływ informacji między poszczególnymi warstwami sieci i zmniejsza złożoność obliczeniową. Relacje między konkretnymi wartościami macierzy dwuwymiarowej (liczba zliczeń w poszczególnych przedziałach) dalej są wychwytywane przez neurony, ponieważ sieć MLP traktuje wszystkie parametry jednakowo, szukając między nimi wzorców, np. korelacji, gdzie liczba zliczeń w jednym przedziale wpływa na zmniejszenie liczby zliczeń w innym. Zatem spłaszczenie rozkładu nie wpłynie na utratę jakiegokolwiek informacji związanej z sąsiedztwem wybranych przedziałów.

Najpierw stworzono wstępny model sieci neuronowych (model bazowy), który następnie poddano modyfikacjom w wyniku jego optymalizacji do wykrywania różnic między cząstkami i antycząstkami. Ten model bazowy składał się z wejścia, które zawierało rozkłady Dalitza o liczbie przedziałów 20×20 (rząd liczby przedziałów wyznaczony w metodzie S_{CP}). Po wypłaszczeniu tego rozkładu otrzymano jednowymiarowy wektor z 400 wartościami. Rozpoczęto od użycia trzech warstw ukrytych, odpowiednio o liczbie neuronów 400, 100 i 32. W każdej warstwie funkcją aktywacją była funkcja oparta na uczeniu pewnych cech dzięki rzadszym aktywacjom tych jednostek (funkcja ReLU). Jako funkcję aktywacji warstwy wyjściowej przyjęto standardowo używaną funkcję sigmoidalną. Każdą ze 100 próbek zawierającą cząstki i antycząstki podzielono na zbiory treningowy i testowy w jednakowej proporcji 50%:50%. W efekcie powstały cztery paczki danych, 2 treningowe (po jednej dla cząstek i antycząstek) i 2 testowe tak samo rozdzielone jak paczki treningowe. Każdy zbiór testowy dodatkowo podzielono na 15 statystycznie mniejszych paczek, aby model miał więcej przypadków do testowania. Ze wzglę-

du na ograniczoną statystykę danych, nie wydzielano dodatkowo zbioru walidacyjnego. Użyto standardowego optymalizatora do iteracji aktualizacji wag sieci opartego na stochastycznym spadku gradientu, który wykorzystuje adaptacyjne oszacowanie pierwszego i drugiego momentu – optymalizatora Adam (z ang. Adaptive Moment Estimation). Do oceny pomyłek klasyfikacji, czyli funkcję straty przyjęto funkcję opartą na obliczeniu logarytmów, czyli funkcję entropii krzyżowej. Tak zdefiniowany model bazowy z parametrami startowymi został poddany optymalizacji na danych z asymetrią. W kolejnych rozdziałach 4.3.3 i 4.3.5 opisano proces doboru parametrów modelu, tak, aby był on jak najbardziej efektywny do klasyfikacji i tym samym do wykrywania różnic między częstkami i antycząstkami.

4.3.3 Dobór liczby przedziałów na rozkładach Dalitza

Przeprowadzono analizę pozwalającą na wyznaczenie najkorzystniejszej liczby przedziałów na rozkładach Dalitza z danymi wejściowymi do modelu. Opisany powyżej model bazowy został wytrenowany 100 razy (raz na każdej ze 100 próbek, tylko zbiorze treningowym), za każdym razem zmieniając liczbę przedziałów na rozkładzie Dalitza. Zakres liczby przedziałów jaki został poddany testowaniu wynosił od 3 do 50. Skuteczność została wyznaczona na podstawie poprawnych przewidywań modelu w zbiorze testowym, na każdej z 15 paczek danych. Wyznaczona ze 100 modeli (100 próbek z asymetrią) średnia skuteczność w zależności od liczby przedziałów na każdej z obu osi rozkładu Dalitza została przedstawiona na rysunku 23 (całkowita liczba przedziałów na rozkładzie Dalitza jest kwadratem liczby testowanych przedziałów). Obserwowane są duże zmiany średniej skuteczności klasyfikacji w zależności od liczby przedziałów. Ta średnia skuteczność jest zatem miarą czułości na wielkość asymetrii między rozkładami cząstek i antycząstek. Zarówno zbyt duża, jak i zbyt mała liczba przedziałów prowadzi do znaczącego zmniejszenia skuteczności modelu. W przypadku zbyt dużej liczby przedziałów istniejąca asymetria między częstkami i antycząstkami może być rozmywana i tym samym różnice między klasami są zacierane, co zmniejsza skuteczność klasyfikacji. Z kolei zbyt mała liczba przedziałów oznacza, że różnice między klasami nie są zbyt wyraziste. Największą skuteczność osiągnięto przy liczbie przedziałów równej siedem na każdej z osi rozkładu Dalitza, co daje 49 przedziałów na całym rozkładzie.



Rysunek 23: Średnia skuteczność modelu wyrażona w procentach w funkcji liczby przedziałów na każdej z obu osi rozkładu Dalitza otrzymana na próbkach z asymetrią. W przypadku dwuwymiarowego rozkładu Dalitza liczba przedziałów jest kwadratem tej wartości.

4.3.4 Dobór statystycznej wielkości zbiorów treningowego i testowego

W idealnym przypadku byłoby, gdyby zbiory treningowy i testowy miały odpowiednio dużą statystykę. Jednakże przy ograniczonej liczbie danych trzeba było znaleźć rozwiązanie najbardziej efektywne. Podobnie jak w przypadku doboru liczby przedziałów, także przy doborze wielkości i liczebności zbiorów treningowych i testowych zdecydowano się na użycie kryterium najwyższej skuteczności przewidywań modelu. To znaczy im wyższą skuteczność przewidywania ma model na próbkach z asymetrią przy określonych parametrach tym rozwiązanie uznawane jest za lepsze. W przypadku zbioru treningowego najcenniejsza okazała się liczba przypadków w paczce, a nie sama liczba paczek. Spowodowane jest to tym, że im liczniejsza paczka tym większa szansa na wykrycie asymetrii między częstkami i antycząstkami. Aby nie tracić informacji zdecydowano się nie rozdzielać zbioru treningowego na mniejsze paczki i na wejściu modelu podać cały zbiór w postaci jednego rozkładu Dalitza z częstkami i drugiego z antycząstkami. Zatem testowanym parametrem w zbiorze treningowym była jedynie liczba rozpadów (częstek lub antycząstek).

Inna sytuacja jest w przypadku zbioru testowego. Tutaj okazało się, że znaczenie ma zarówno liczba paczek, jak i statystyka w każdej z nich. Do testowania modelu, podobnie jak

w przypadku jego trenowania, można byłoby użyć całej statystyki dostępnej w zbiorze, ale w tej sytuacji wynik nie byłby miarodajny, gdyż model przewidziałby tylko jeden raz czy testowany zbiór jest rozkładem cząstek czy antycząstek. W takiej sytuacji średnia skuteczność ze 100 modeli byłaby najwyższa. Dlatego zdecydowano się na podzielenie zbioru testowego na paczki, gdyż daje to możliwość kilkukrotnego sprawdzenia jak dobrze radzi sobie model z klasyfikacją przypadków. Takie podejście pozwala bardziej realistycznie zoptymalizować model pod względem wykrywania różnic między cząstkami i antycząstkami. Przykładowo, gdyby ustalić liczbę paczek w zbiorze testowym na jeden, to model miałby dwie wartości do przewidzenia (cząstka lub antycząstka). Mógłby trafić obie, nie trafić żadnej lub trafić po jednej. W sytuacji bez asymetrii między cząstkami i antycząstkami prawdopodobieństwo wystąpienia każdej z tych trzy sytuacji byłoby rzędu 33%. Zatem nawet jeśli skuteczność przewidywania wynosiłaby 100%, istniałaby duża szansa, że asymetrii w próbce nie ma. Także czułość wykrywania asymetrii nie poprawiłaby się, a dodatkowo pojawiłby się czynnik mogący fałszywie wskazywać asymetrię. Zdecydowano się zatem, że minimalną liczbą paczek w każdym zbiorze testowym (ze 100) będzie 10 (10 paczek z cząstkami i 10 paczek z antycząstkami). Do testowania poddano także całkowitą licznosc zbioru testowego. Przyjęto zakres od 1 000 do 75 000 cząstek i tyle samo antycząstek (całkowita liczba przypadków w każdej próbce wynosi 100 000 cząstek i 100 000 antycząstek i są one rozdzielane między zbiory treningowy i testowy). W przypadku zbioru treningowego sprawdzono jego podział na liczbę paczek w zakresie od 10 do 100.

Na rysunku 24 przedstawiono średnią skuteczność przewidywań modelu w funkcji liczby paczek na jakie podzielono zbiory testowe w zależności od całkowitej liczebności zbioru treningowego otrzymaną na próbkach z asymetrią. Z kolei na rysunku 25 przedstawiono średnią skuteczność modelu w funkcji liczebności zbioru treningowego w zależności od liczby paczek na jakie podzielono zbiór testowy.

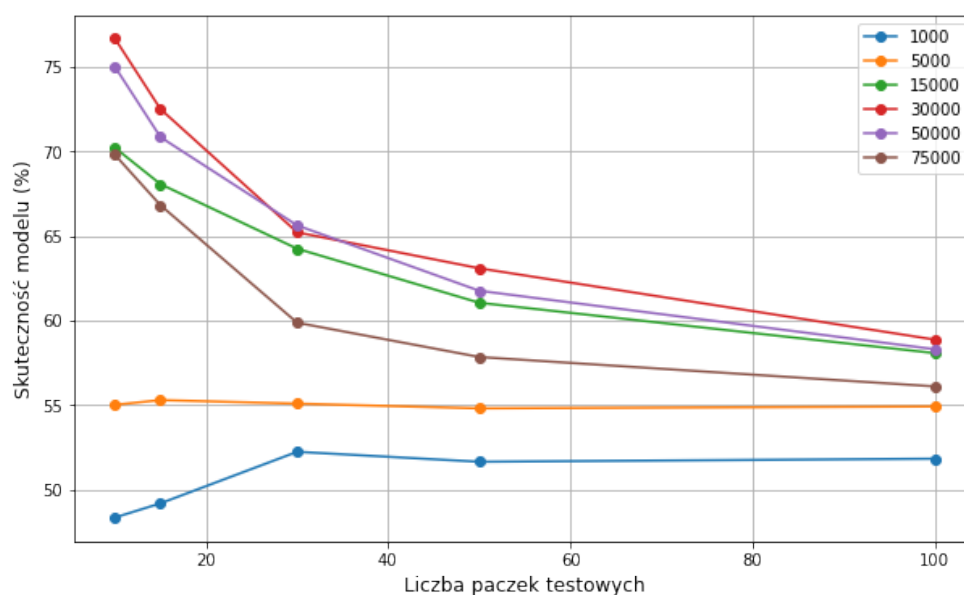
Zaobserwowano, że mniejsza liczba paczek na jakie dzielono zbiory testowe poprawia skuteczność klasyfikacji. Jak należało się spodziewać, także większa liczba zliczeń w zbiorze treningowy poprawia skuteczność modelu, ale w pewnym momencie następuje odwrócenie sytuacji i dalsze zwiększanie statystyki zmniejsza już skuteczność. Jest spowodowane tym, że zwiększając statystykę zbioru treningowego zmniejsza się statystykę zbioru testowego i odwrotnie. Przy danej statystyce zliczeń (łącznie 200 000 cząstek i antycząstek w każdej ze 100 próbek) maksimum skuteczności modelu jest osiągnięte przy 30 000 cząstek i tyle samo antycząstek w zbiorze treningowym. Do ostatecznego modelu zdecydowano się zatem na przyjęcie zbioru

treningowego w postaci jednej paczki z 30 000 cząstek i tyle samo antycząstek (co stanowi rzędu $1/3$ dostępnej statystyki danych).

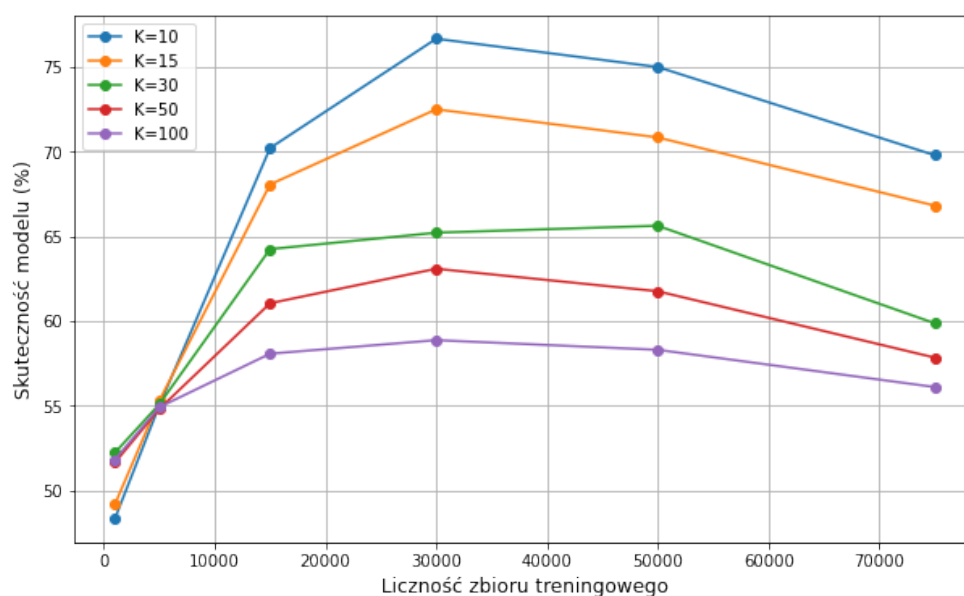
Jeśli zaś chodzi o liczbę paczek na jaką należy podzielić zbiór testowy, to problem jest bardziej skomplikowany. Wydawałoby się, że najlepszym rozwiązaniem byłoby przyjęcie liczby paczek jako 10 (maksimum na rysunku 25), która jest najmniejszą testowaną liczbą. Jak już wspomniano, największej efektywności należałoby oczekiwać, gdyby nie dzielić zbioru testowego na paczki. Jednakże w takiej sytuacji brak asymetrii między cząstkami i antycząstkami mógłby być fałszywie interpretowany jako sygnał takiej asymetrii, gdyż skuteczność przewidywań modelu byłaby wysoka i różna od losowej, która powinna być w okolicach 50% w tej sytuacji. Zauważono zależność odchylenia standardowego skuteczności (tabela 10) w zależności od liczby paczek w zbiorze testowym (K). Pojawia się różnica około jednego punktu procentowego między wartościami $K=10$ i $K=12$, co przy ich podobnych średnich jest argumentem za użyciem $K=12$. Z kolei już między $K=12$ oraz $K=15$ różnica jest tylko 0,1 punktu procentowego. Dlatego nie zwiększano dalej liczby paczek. Rozkłady średniej skuteczności są przedstawione i dyskutowane w szczegółach w dalszej części pracy.

Ostatecznie zdecydowano, że każda próbka danych zostanie podzielona w następujący sposób:

- Zbiór treningowy o statystyce 30 000 cząstek i 30 000 antycząstek bez dalszego podziału na paczki.
- Zbiór testowy o uzupełniającej statystyce 70 000 cząstek i 70 000 antycząstek podzielony na 12 paczek po 5 833 cząstek i tyle samo antycząstek w każdej.



Rysunek 24: Średnia skuteczność przewidywań uzyskana na 100 próbkach z asymetrią między cząstkami i antycząstkami w funkcji liczby paczek na jakie podzielono zbiory testowe w zależności od całkowitej liczebności zbioru treningowego od 1 000 do 75 000 cząstek i tyle samo antycząstek (różne kolory linii).



Rysunek 25: Średnia skuteczność przewidywań modelu uzyskana na 100 próbkach z asymetrią między cząstkami i antycząstkami w funkcji liczebności zbioru treningowego (łączna statystyka cząstek i antycząstek jest sumą podanej liczby) przy różnych liczbach paczek na jakie podzielono zbiory testowe (K).

Tabela 10: Odchylenie standardowe średniej skuteczności (σ) ze 100 próbek z asymetrią między cząstkami i antycząstkami przy liczbie paczek w zbiorze testowym (K) wynoszącej 10, 12 i 15.

K	σ (%)
10	11,12
12	10,01
15	9,91

4.3.5 Dobór parametrów sieci neuronowej

Po ustaleniu liczby przedziałów oraz odpowiedniej proporcji zbiorów treningowych i testowych pozostały do wyznaczenia parametry charakteryzujące sieć neuronową. Testowano liczbę warstw ukrytych i typ ich funkcji aktywacji. Jak się okazało trzy warstwy ukryte z liczbami neuronów odpowiednio 400, 100, 32 przyjęte w modelu bazowym były wystarczające i nie wymagały zmian.

Zmniejszenie liczby przedziałów na rozkładach Dalitza z 20x20 do 7x7, czyli z 400 parametrów zejście do 49 spowodowało, że jedna warstwa ukryta z liczbą neuronów w okolicy liczby parametrów wejściowych jest w stanie osiągać wyniki w granicach 66% średniej skuteczności modelu w próbkach z asymetrią między cząstkami i antycząstkami. Po dodaniu drugiej warstwy ukrytej z 10-20 neuronami wynik poprawiał się do około 73%. Kolejne warstwy czy też zwiększanie liczby neuronów nie były w stanie poprawić już tego wyniku.

Jeśli zaś chodzi o wybór funkcji aktywacji, to model wykazał się stabilnością i niezależnie od wybranej funkcji (testowano funkcję ReLU, sigmoidalną, eksponencjalną, liniową oraz tangens) lub kombinacji kilku z nich w różnych warstwach ukrytych wynik był stabilny i oscylował wokół tej samej wartości 73%.

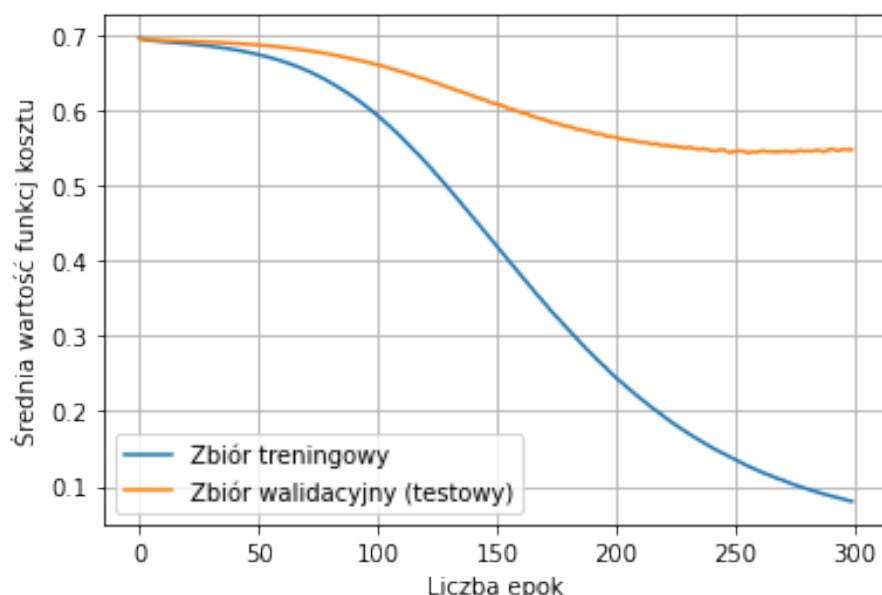
Kolejnym testowanym parametrem była liczba epok, czyli ile razy modelowi pokazano dane na etapie trenowania. Sprawdzano czy model nie dopasowuje się nadmiernie do zbioru treningowego. Efekt ten można zauważyć, gdy funkcja kosztu rośnie na zbiorze walidacyjnym. Jako, że statystyka danych była ograniczona, to wydzielenie zbioru walidacyjnego wpłynęłoby na pogorszenie efektywności. Dlatego sprawdzono jak zachowują się dane treningowe i testowe w kolejnych epokach. Wykorzystano tutaj średnią arytmetyczną funkcji kosztu ze wszystkich

próbek, a wyniki zostały przedstawione na rysunku 26. Do obliczania funkcji straty wykorzystano stosowaną w klasyfikacji binarnej logarytmiczną funkcję straty ($loss$) zdefiniowaną w następujący sposób:

$$loss = -(y_r * \log(y_p) + (1 - y_r) * \log(1 - y_p)), \quad (11)$$

gdzie: y_r jest rzeczywistą (prawdziwą) wartością wyjścia, a y_p jest wartością przewidywaną.

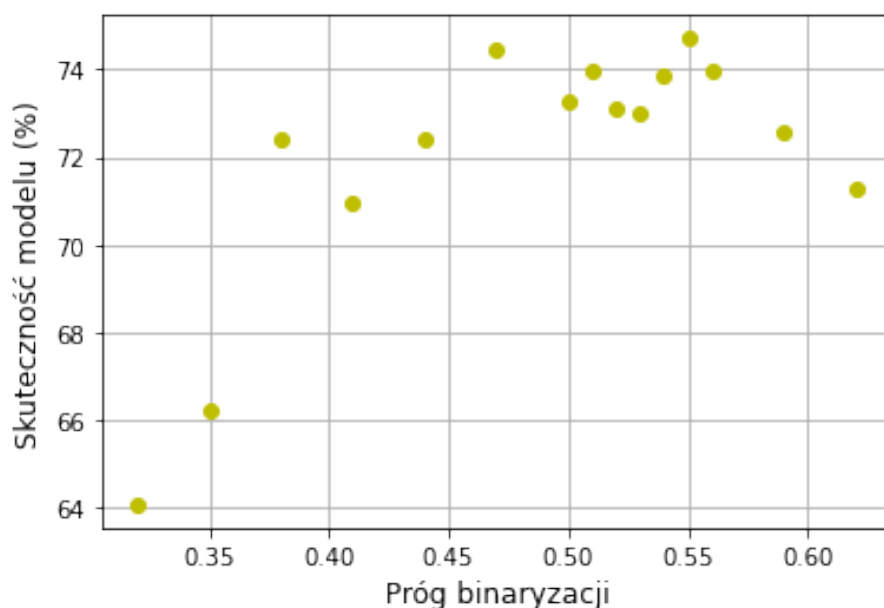
W testowanej statystyce danych (rysunek 26), około setnej epoki obserwuje się zmniejszanie funkcji kosztu zarówno w zbiorze treningowym jak i testowym. Przy około 200 epokach wartość funkcji kosztu w zbiorze testowym zaczyna się ustalać na stałym poziomie. Natomiast w zbiorze treningowym funkcja kosztu zmniejsza się dalej. Do trenowania modelu przyjęto 200 epok.



Rysunek 26: Średnia wartość funkcji kosztu ze 100 próbek z asymetrią między częstkami i antycząstkami w zależności od liczby epok wyznaczona na zbiorach treningowym (linia niebieska) i testowym (linia pomarańczowa). W celu zachowania analogii do klasycznego postępowania z sieciami neuronowymi zbiór testowy na rysunku został dodatkowo nazwany zbiorem walidacyjnym.

Jako, że model ma przewidywać binarne wartości 0 lub 1, czyli czy dana paczka danych należy do częstek (0), czy antycząstek (1), to na wyjściu przyjęto funkcję sigmoidalną przyjmującą wartości z zakresu od 0 do 1. W celu wyznaczenia progu binaryzacji obliczono średnie

wartości skuteczności modelu ze 100 próbek z asymetrią między cząstkami i antycząstkami przy różnych progach binaryzacji z zakresu 0,32-0,62 (rysunek 27). Od około wartości 0,45 progu binaryzacji do około 0,55 obserwowana jest podobna skuteczność, która jest jednocześnie najwyższa w całym badanym zakresie. Nie da się jednoznacznie wyznaczyć wartości maksimum średniej skuteczności, dlatego przyjęto za próg binaryzacji wartość 0,50.

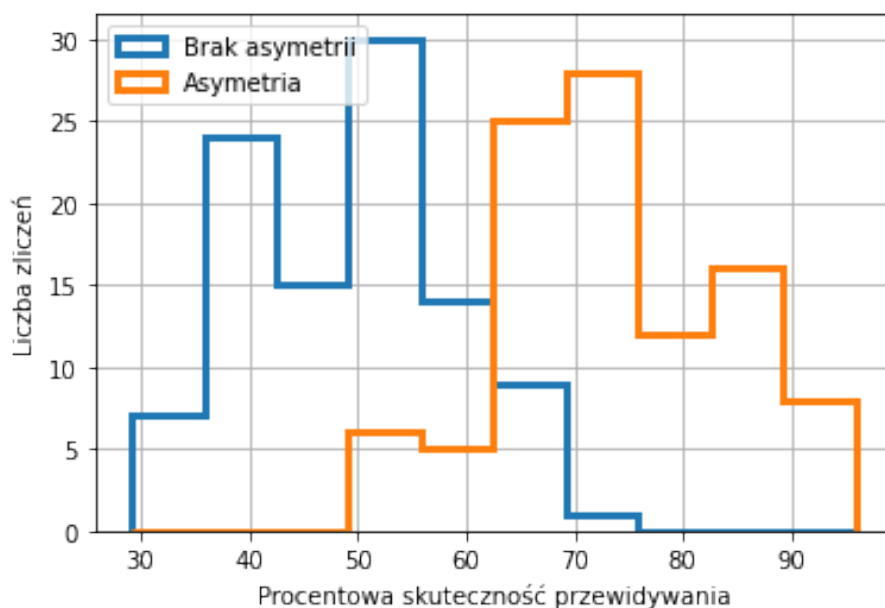


Rysunek 27: Średnia skuteczność przewidywania modelu ze 100 próbek z asymetrią między cząstkami i antycząstkami w zależności od wartości progu binaryzacji.

Wszystkie ustalone powyżej parametry przyjęto jako ostateczny model sieci neuronowej i użyto do analizy 100 próbek danych z asymetrią między cząstkami i antycząstkami, jak i bez tej asymetrii do ostatecznego oszacowania czułości metody.

4.3.6 Czułość modelu

Na rysunku 28 przedstawiono rozkłady skuteczności ostatecznego modelu sieci neuronowych otrzymane na 100 próbkach z asymetrią między cząstkami i antycząstkami oraz bez asymetrii. Rozkłady te różnią się od siebie. W sytuacji gdy występuję niezerowa asymetria między cząstkami i antycząstkami, to rozkład skuteczności przewidywać jest przesunięty w kierunku wyższych wartości. Natomiast w przypadku braku asymetrii między cząstkami i antycząstkami rozkład skuteczności jest symetryczny względem wartości 50%.



Rysunek 28: Skuteczność przewidywań modelu sieci neuronowych w procentach mierzona w 100 próbkach z asymetrią między cząstkami i antycząstkami (linia pomarańczowa) i bez tej asymetrii (linia niebieska).

W tabeli 11 podano średnią skuteczność modelu wyznaczoną ze 100 próbek z i bez asymetrii między cząstkami i antycząstkami (podano także wartości minimalną i maksymalną oraz odchylenie standardowe średniej). Średnia skuteczność modelu jest zdecydowanie wyższa w przypadku niezerowego łamania parzystości CP i wynosi $73,4 \pm 10,8\%$ niż w przypadku braku łamania CP , gdzie wynosi $49,1 \pm 9,4\%$ (zgodne z oczekiwaną wartością 50%). Różnice także są widoczne w wartościach minimalnej i maksymalnej skuteczności. W przypadku próbek z asymetrią wartość minimalna jest niższa niż maksymalna wartość mierzona w próbkach bez asymetrii. Oznacza to, że w zakresie 50-75% nie da się jednoznacznie stwierdzić czy istnieje różnica między cząstkami i antycząstkami, czy też jej nie ma. Jednakże większa statystyka danych wpłynie na skuteczność modelu. Im większa statystyka tym większe różnice w efektywności modelu między próbkami z i bez asymetrii.

Tabela 11: Średnia skuteczność modelu sieci neuronowych wyznaczona ze 100 próbek, jej wartość maksymalna i minimalna oraz odchylenie standardowe skuteczności (σ) w sytuacji braku oraz występowania łamania parzystości CP .

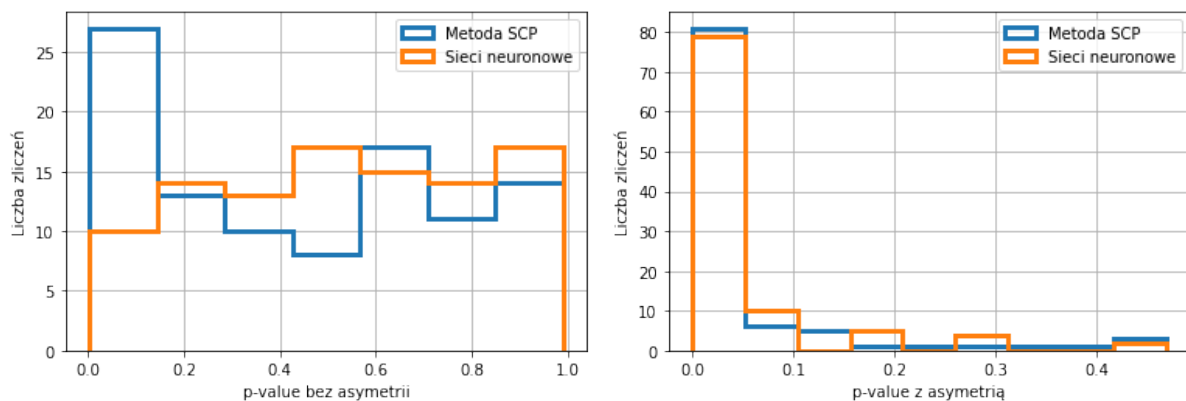
Łamanie CP	Średnia skuteczność (%)	Maksimum (%)	Minimum (%)	σ (%)
Brak	49,1	75,0	29,2	9,4
Asymetria	73,4	95,8	50,0	10,8

Sprawdzono także macierz pomyłek modelu, a wyniki przedstawiono w tabeli 12. Niezależnie czy mamy do czynienia z cząstkami czy antycząstkami model jest tak samo skuteczny w ich przewidywaniu. Liczby na diagonalu są rzędu dziewięć, co oznacza dziewięć poprawnych przewidywań modelu na 12 prób (paczek w zbiorze testowym). Daje to efektywność rzędu 75%. Jeśli zaś chodzi o liczbę pomyłek, to jest ona rzędu trzech na 12 paczek, co stanowi pozostałe 25%. Pokazuje to, że moc metody poszukiwania różnic między cząstkami i antycząstkami opartej na sieciach neuronowych jest rzędu 75%. Ta wielkość jednakże nie mówi na jakim poziomie statystycznym potwierdzone są różnice. W tym celu zbudowano test statystyczny pozwalający na wyznaczenie prawdopodobieństwa zgodności cząstek i antycząstek. Jako hipotezę zerową przyjęto zgodność rozkładów cząstek i antycząstek, czyli odpowiadający temu rozkład skuteczności modelu z rysunku 28 (rozkład niebieski). Aby było możliwe też wytestowanie czy metoda nie zwraca fałszywych sygnałów różnic między cząstkami i antycząstkami tak, gdzie ich nie ma, to za każdym razem (w każdej próbce) wyznaczano rozkład hipotezy zerowej z losowego przypisywania klasy przypadkowi jako cząstka lub antycząstka. Następnie korzystając z testu χ^2 wyznaczono prawdopodobieństwo zgodności z tym rozkładem średniej skuteczności mierzonej w każdej ze 100 próbek, zarówno z asymetrią jak i bez asymetrii. Wyniki zostały przedstawione na rysunku 29. Otrzymane wartości zostały porównane względem metody, którą uznano za referencyjną, czyli względem metody S_{CP} .

W sytuacji braku różnic między cząstkami i antycząstkami metoda oparta na modelu sieci neuronowych nie generuje fałszywych sygnałów różnic. Rozkład prawdopodobieństwa jest płaski (pomarańczowy histogram z lewego rysunku 29). Jeśli zaś chodzi o próbki z asymetrią (prawy rysunek 29), to model sieci neuronowych potwierdza istnienie w próbkach różnic między cząstkami i antycząstkami. Rozkład prawdopodobieństwa w tym przypadku nie jest rozkładem płaskim i dodatkowo jest on podobny do otrzymanego z metody S_{CP} .

Tabela 12: Macierz pomyłek ilustrująca ile cząstek i antycząstek zostało poprawnie przypisanych ze zbioru testowego, a także w ilu przypadkach model się pomylił. Wyciągnięto średnią ze 100 modeli.

Przewidywane \ Rzeczywiste	Cząstki	Antycząstki
	Cząstki	Antycząstki
Cząstki	9,0	3,0
Antycząstki	3,4	8,6



Rysunek 29: Prawdopodobieństwo (p-value) zgodności średniej skuteczności w próbce z losowanym rozkładem zakładającym brak różnic między cząstkami i antycząstkami. Po lewej stronie wyniki otrzymane na danych z brakiem asymetrii a po prawej stronie otrzymane na danych z asymetrią. Wyniki z modelu sieci neuronowych (linia pomarańczowa) porównane z wartościami z metody S_{CP} (linia niebieska).

W tabeli 13 porównano odpowiedzi metody S_{CP} i modelu sieci neuronowych w każdej ze 100 próbek danych z asymetrią między cząstkami i antycząstkami i bez tej asymetrii. Podano w ilu próbkach zmierzone prawdopodobieństwo zgodności z brakiem różnic między cząstkami i antycząstkami przekracza trzy i pięć standardowych odchyłeń. Odpowiedzi obu metod są porównywalne. Metody nie potwierdzają różnic tam gdzie ich nie ma i potwierdzają różnice jeśli one występują. To potwierdzenie różnic jest na podobnym poziomie w obu metodach. Czułość metody S_{CP} na odchylenie powyżej 3σ wynosi 31%, a czułość sieci neuronowych jest nieznacznie wyższa i wynosi 36%.

Tabela 13: Liczba próbek na 100, w których prawdopodobieństwo zgodności rozkładów cząstek i antycząstek jest powyżej trzech i pięciu standardowych odchyłeń (σ) po zastosowaniu metody S_{CP} i modelu sieci neuronowych w sytuacji braku łamania parzystości CP i jego występowania.

Łamanie CP	Sieci Neuronowe	Odchylenia powyżej 3σ	Odchylenia powyżej 5σ
Brak	Nie	0	0
Brak	Tak	0	0
Asymetria	Nie	31	0
Asymetria	Tak	36	0

Mimo że różnica między cząstkami i antycząstkami nie jest duża (5% łamania CP w amplitudach rezonansu K^* oznacza, że asymetrie między cząstkami i antycząstkami w całym rozkładzie Dalitza są rzędu promila), sieci neuronowe dały obiecujące wyniki pozwalające stwierdzić, że są one dobrym narzędziem do poszukiwania różnic między próbkami. Te różnice były badane na przykładzie łamania parzystości CP . Jednakże mogą to być dowolne różnice między dowolnymi porównywanymi próbkami. Zamiast rozkładów cząstek i antycząstek mogą to być dwa zdjęcia z tomografu płuc pacjenta. Jedno wykonane w jakimś okresie czasu, a drugie np. rok później. Potwierdzenie różnic między tymi zdjęciami na poziomie promila może być pomocne w diagnozowaniu różnego rodzaju chorób.

Powyższe badania pokazują, że warto też rozważyć inne algorytmy uczenia maszynowego do poszukiwania różnic między badanymi próbkami, jak innych typów sieci neuronowych (konwolucyjnych, zespolonych, grafowych itp.), jak również warto się zastanowić nad użyciem metod nienadzorowanego uczenia maszynowego, jak np. klasteryzacji.

5 Podsumowanie

Celem projektu było opracowanie algorytmu potrafiącego potwierdzać różnice między porównywanymi próbkami. Wykonano to na przykładzie danych fizyki cząstek elementarnych z łamaniem parzystości CP w trzyciałowych rozpadach barionów powabnych.

Jako metodę bazową użyto metodę opartą na podziale przestrzeni kinematycznej na przedziały, używaną w eksperymencie LHCb. Podjęto się wyzwania opracowania skuteczniejszej metody rozpoznawania różnic między rozkładami cząstek i antycząstek. W tym celu najpierw zaproponowano użycie algorytmu Imana-Conovera do poprawy czułości metody bazowej. Jednakże wyniki nie były jednoznaczne. Oczekiwano, że algorytm powinien poprawić czułość metody bazowej z uwagi na swoją konstrukcję, co niestety się nie udało. Mimo to udało się wykazać, że algorytm ten nie generuje fałszywych sygnałów różnic między porównywanymi próbkami przy względnie dużej statystyce danych, jak 200 000. Zauważono, że w sytuacji znacznie mniejszych statystyk algorytm Imana-Conovera może jednak wykazywać poprawę czułości metody bazowej. Ma to istotne znaczenie, gdyż w fizyce cząstek elementarnych często statystyki danych są bardzo małe, rzędu kilku zliczeń, jak w przypadku bardzo rzadkich rozpadów różnych cząstek z przekrojami czynnymi mniejszymi niż 10^{-10} . Ponadto zjawiska fizyczne nie zawsze występują w całym obszarze przestrzeni fazowej. Mogą być lokalne, jak ma to miejsce z łamaniem parzystości CP . Dlatego można się zastanowić czy algorytm Imana-Conovera nie będzie pomocny do wykrycia tego efektu w jakimś obszarze przestrzeni fazowej, gdzie statystyka danych jest bardzo mała. Sam fakt, że algorytm nie ma wykrywanej fałszywej asymetrii i potwierdza tam gdzie ona jest, pozwala bezpiecznie korzystać z niego korzystać. Co do samego algorytmu poszerzył on przestrzeń rozwiązań zwiększając liczbę badanych przedziałów biorących.

Wobec braku powodzenia z zastosowaniem algorytmu Imana-Conovera podjęto się całkowicie nowego wyzwania polegającego na użyciu modelu sieci neuronowych do poszukiwania łamania parzystości CP . Badania na danych testowych pokazały, że ta metoda może osiągnąć skuteczność rzędu 75% przy statystyce danych wynoszącej 200 000 zliczeń. Większe statystyki danych będą wpływały na zwiększenie skuteczności tej metody. Zauważono też kilku procentowy wzrost wykrywalności różnic między porównywanymi próbkami względem metody bazowej. Miarą tych porównań był pomiar prawdopodobieństwa zgodności rozkładów przy założeniu braku różnic między nimi. Wzrost czułości metody opartej na sieciach neuronowych na występujące różnice między próbkami nie jest jednoznaczny i wymaga dalszych testów. Pokazano jednak, że metody oparte na uczeniu maszynowym mogą być skutecznie używane do

poszukiwań łamania parzystości *CP*, co nigdy wcześniej nie było rozważane. Tym samym tego typu podejścia mogą być zastosowane do poszukiwania jakichkolwiek różnic między porównywanymi próbkami, jak np. zdjęć pacjentów uzyskiwanych w tomografach komputerowych do poszukiwania zmian chorobowych.

Przeprowadzone testy raz jeszcze potwierdzają, jak istotną dziedziną jest szeroko rozumiana sztuczna inteligencja. Mimo ograniczenia się do sieci neuronowych, następnie do jej konkretnej implementacji (MLP), a także specyficznego przetwarzania danych, kombinacje parametrów do przetestowania mnożyły się z każdym krokiem. Ważnym wnioskiem, który można wyciągnąć z tych badań jest to, że warto rozwijać algorytmy oparte na sieciach neuronowych do poszukiwania bardzo małych różnic na poziomie promila. Możliwości do poprawy jest dużo, jak na przykład użycie bardziej skomplikowanych typów sieci neuronowych, czy też rozważenie użycia nienadzorowanych metod uczenia maszynowego.

W pracy magisterskiej próbowano różnych podejść rozwiązania tego samego problemu celem znalezienia najbardziej optymalnego sposobu badania danych. Temat jest obszerny, każda z metod dała obiecujące wyniki, co sprawia, że warto je dalej w tym temacie rozwijać.

Literatura

1. Christenson, J., Cronin, J., Fitch, V. & Turlay, R. Evidence for the 2π decay of the K_2^0 meson. *Phys. Rev. Lett.* **13**, 138–140. doi:10.1103/PhysRevLett.13.138 (1964).
2. Aubert, B. *et al.* Observation of CP violation in the B^0 meson system. *Phys. Rev. Lett.* **87**, 091801. doi:10.1103/PhysRevLett.87.091801. arXiv: hep-ex/0107013 [hep-ex] (2001).
3. Abe, K. *et al.* Observation of large CP violation in the neutral B meson system. *Phys. Rev. Lett.* **87**, 091802. doi:10.1103/PhysRevLett.87.091802. arXiv: hep-ex/0107061 [hep-ex] (2001).
4. Aubert, B. *et al.* Observation of direct CP violation in $B^0 \rightarrow K^+\pi^-$ decays. *Phys. Rev. Lett.* **93**, 131801. doi:10.1103/PhysRevLett.93.131801. arXiv: hep-ex/0407057 [hep-ex] (2004).
5. Chao, Y. *et al.* Evidence for direct CP violation in $B^0 \rightarrow K^+\pi^-$ decays. *Phys. Rev. Lett.* **93**, 191802. doi:10.1103/PhysRevLett.93.191802. arXiv: hep-ex/0408100 [hep-ex] (2004).
6. Aaij, R. *et al.* Observation of CP violation in $B^\pm \rightarrow DK^\pm$ decays. *Phys. Lett.* **B712**, 203. doi:10.1016/j.physletb.2012.04.060. arXiv: 1203.3662 [hep-ex] (2012).
7. Aaij, R. *et al.* First observation of CP violation in the decays of B_s^0 mesons. *Phys. Rev. Lett.* **110**, 221601. doi:10.1103/PhysRevLett.110.221601. arXiv: 1304.6173 [hep-ex] (2013).
8. Aaij, R. *et al.* Observation of CP violation in charm decays. *Phys. Rev. Lett.* **122**, 211803. doi:10.1103/PhysRevLett.122.211803. arXiv: 1903.08726 [hep-ex] (2019).
9. Aaij, R. *et al.* A measurement of the CP asymmetry difference in $\Lambda_c^+ \rightarrow pK^-K^+$ and $p\pi^-\pi^+$ decays. *JHEP* **03**, 182. doi:10.1007/JHEP03(2018)182. arXiv: 1712.07051 [hep-ex] (2018).
10. Aaij, R. *et al.* Measurement of matter-antimatter differences in beauty baryon decays. *Nature Phys.* **13**, 391–396. doi:10.1038/nphys4021. arXiv: 1609.05216 [hep-ex] (2017).

11. Aaij, R. *et al.* Search for CP violation and observation of P violation in $\Lambda_b^0 \rightarrow p\pi^-\pi^+\pi^-$ decays. *Phys. Rev. D* **102**, 051101. doi:10.1103/PhysRevD.102.051101. arXiv: 1912.10741 [hep-ex] (2020).
12. Aaij, R. *et al.* Search for CP violation in $\Xi_c^+ \rightarrow pK^-\pi^+$ decays using model-independent techniques. *Eur. Phys. J. C* **80**, 986. doi:10.1140/epjc/s10052-020-8365-0. arXiv: 2006.03145 [hep-ex] (2020).
13. Ukleja, A. Poszukiwania sygnałów nowej fizyki i źródeł łamania parzystości CP w wielociałowych rozpadach cząstek powabnych w eksperymencie LHCb. ISBN: **978-83-66364-99-8**, Wydawnictwa AGH, Kraków. <https://www.wydawnictwo.agh.edu.pl/produkt/112-poszukiwania-sygnaLOW-nowej-fizyki-i-zrodel-lamania-parzystosci-cp-w-wielocialowych> (2021).
14. Chala, M., Lenz, A., Rusov, A. V. & Scholtz, J. ΔA_{CP} within the Standard Model and beyond. *JHEP* **07**, 161. doi:10.1007/JHEP07(2019)161. arXiv: 1903.10490 [hep-ph] (2019).
15. Bianco, S., Fabbri, F. L., Benson, D. & Bigi, I. A Cicerone for the physics of charm. *Riv. Nuovo Cim.* **26**, 1–200. doi:10.1393/ncr/i2003-10003-1. arXiv: hep-ex/0309021 (2003).
16. Bigi, I. I. Probing CP asymmetries in charm baryons decays. arXiv: 1206.4554 [hep-ph] (June 2012).
17. Aaij, R. *et al.* Observation of CP violation in $B^\pm \rightarrow DK^\pm$ decays. *Phys. Lett. B* **712**. [Erratum: Phys.Lett.B 713, 351 (2012)], 203–212. doi:10.1016/j.physletb.2012.04.060. arXiv: 1203.3662 [hep-ex] (2012).
18. Aaij, R. *et al.* First observation of CP violation in the decays of B_s^0 mesons. *Phys. Rev. Lett.* **110**, 221601. doi:10.1103/PhysRevLett.110.221601. arXiv: 1304.6173 [hep-ex] (2013).
19. Aaij, R. *et al.* Direct CP violation in charmless three-body decays of B^\pm mesons. *Phys. Rev. D* **108**, 012008. doi:10.1103/PhysRevD.108.012008. arXiv: 2206.07622 [hep-ex] (2023).
20. Alves Jr., A. A. *et al.* The LHCb detector at the LHC. *JINST* **3**, S08005. doi:10.1088/1748-0221/3/08/S08005 (2008).

21. <https://www.tensorflow.org>.
22. https://en.wikipedia.org/wiki/Simple_random_sample.
23. Iman, R. & Conover, W. A distribution-free approach to inducing rank correlation among input variables. <https://www.uio.no/studier/emner/matnat/math/STK4400/v05/undervisningsmateriale/A%20distribution-free%20approach%20to%20rank%20correlation.pdf> (1982).
24. <https://blogs.sas.com/content/iml/2012/02/08/use-the-cholesky-transformation-to-correlate-and-uncorrelate-variables.html>.
25. Justyna, D. Generator Monte Carlo wiązki terapeutycznej w oparciu o charakterystykę przestrzeni fazowej IAEA (2023).
26. <https://blogs.sas.com/content/iml/2021/06/16/geometry-iman-conover-transformation.html>.
27. Jastrzębski, T. Zastosowanie symulacji Monte Carlo z korektą Imana-Conovera w pomiarze ryzyka systemowego. https://wzr.ug.edu.pl/.zif/5_10.pdf (2017).
28. <https://www.sztuczna inteligencja.org.pl/kurs/sztuczna-inteligencja-dla-poczatkujacych/sztuczne-sieci-neuronowe/150162-2/>.
29. https://kcir.pwr.edu.pl/~witold/ai/ml_nndeep_s.pdf.
30. Cardon, D., Cointet, J.-P. & Mazieres, A. Neurons spike back. The invention of inductive machines and the artificial intelligence controversy. *ISBN: 978-23-48040-68-9*, Reseaux 2018/5 (n. 211), pp. 173–220. doi:10.3917/res.211.0173 (2018).
31. <https://home.agh.edu.pl/~horzyk/lectures/miw/SI-MIW-SztuczneSieciNeuronowe.pdf>.
32. https://en.wikipedia.org/wiki/Sturges%27s_rule.
33. <https://www.python.org>.
34. <https://eitca.org/artificial-intelligence/eitc-ai-dlpp-deep-learning-with-python-and-pytorch/neural-network/building-neural-network/examination-review-building-neural-network/why-do-we-need-to-flatten-images-before-passing-them-through-the-network/>.