



Dr hab. Justyna Łagoda
Narodowe Centrum Badań Jądrowych
Departament Badań Podstawowych

Warszawa, 15.06.2020 r.

Recenzja rozprawy doktorskiej mgr. Michała Olszewskiego zatytułowanej „Machine learning approach to $H \rightarrow \tau\tau$ analysis in the CMS experiment”

Mgr Michał Olszewski jest członkiem międzynarodowej współpracy CMS, prowadzącej badania przy akceleratorze LHC w CERN. Jednym z głównych zagadnień podejmowanych przez CMS jest badanie bozonu Higgosa. Praca doktorska magistra Olszewskiego pt. „Machine learning approach to $H \rightarrow \tau\tau$ analysis in the CMS experiment” przedstawia metody poszukiwań bozonu Higgosa, w kanale rozpadu na leptonów τ , gdzie jeden z leptonów τ rozpada się następnie na mion i neutrino, a drugi semileptonowo. Pan Olszewski odtworzył najpierw oficjalną metodę analizy używaną w zespole CMS, a następnie opracował własną metodę klasyfikacji, opartą o uczenie maszynowe, i pokazał, jak jej wprowadzenie może polepszyć wyniki uzyskiwane przez zespół.

Rozpad bozonu Higgosa na parę leptonów τ jest interesujący z wielu względów. Jest to drugi najczęstszy – po rozpadzie na parę kwarków b – rozpad na fermiony, łatwiejszy do zaobserwowania ze względu na mniejsze tło, co pozwala na użycie go do pomiaru sprzężenia bozonu Higgosa do fermionów, badania parzystości CP bozonu Higgosa lub też do poszukiwań dodatkowych bozonów Higgosa, które przewidywane są przez rozszerzenia Modelu Standardowego.

Z drugiej jednak strony, poszukiwanie sygnału tego rozpadu jest trudnym zadaniem ze względu na niemożność zaobserwowania bezpośrednio krótkożyciowego leptonu τ i liczne kanały jego rozpadu (w których zawsze część energii unoszona jest przez neutrino). Dlatego też tak istotne jest poszukiwanie skutecznych metod odróżniania sygnału od tła (np. pochodzącego z rozpadu $Z \rightarrow \tau\tau$). Tradycyjnym podejściem jest stosowanie cięć na wybrane zmienne kinematyczne przypadków (lub ich kombinacje), często połączone – jak w opisywanej w pracy analizie – z podziałem przypadków na pewne kategorie zdefiniowane przez stany końcowe, które analizowane są oddzielnie w celu zmaksymalizowania efektywności selekcji i czystości otrzymywanych próbek. Zagadnieniu poszukiwania ulepszeń takiej selekcji poświęcona jest rozprawa mgr. Olszewskiego.

Praca została napisana w języku angielskim i składa się ze wstępu, sześciu rozdziałów oraz podsumowania, uzupełnionych o pięć dodatków. Omówię teraz kolejne rozdziały, uwzględniając moje do nich uwagi.

We wstępie autor przedstawia krótko tematykę i cel pracy, oraz jej organizację.

Rozdział I opisuje teoretyczne podstawy Modelu Standardowego oraz mechanizm spontanicznego łamania symetrii. Przypomina również właściwości bozonu Higgosa, takie jak masa, sprzężenia oraz kanały produkcji w zderzeniach proton-proton. Autor omawia również

kanały rozpadu bozonu Higgsa, krótko wspominając metody doświadczalne używane do identyfikacji tych rozpadów.

Rozdział ten zawiera również informacje o leptonie τ , takie jak jego czas życia, zasięg i najważniejsze kanały rozpadu. Na zakończenie pokazany jest rozkład zmiennej zwanej „hadronic tau mass” lub „visual mass”, która jest jak przypuszczam sumą energii cząstek widocznych w detektorze – prawdopodobnie tych, które zostały uznane za pochodzące z rozpadu leptonu τ . Opis tego rysunku (1.9) jest bardzo skrótowy i mało precyzyjny.

Rozdział 2 autor rozpoczyna od przypomnienia historii ośrodka CERN, budowy zderzacza LHC i detektora CMS. Przedstawia następnie szczegółowo parametry akceleratora wraz z definicją świetlności i informacją o ilości zebranych danych. Dalej opisana jest budowa detektora CMS, gdzie omówione są po kolei jego podzespoły. Miejscami autor przytacza tu duże ilości szczegółów technicznych, które jakkolwiek ciekawe, nie zawsze są istotne dla jego analizy, a ich natłok powoduje, że trudno wyłowić te ważniejsze. Rozdział kończy się opisem systemu wyzwalania stosowanym w eksperymencie.

Kolejny rozdział poświęcony jest opisowi rekonstrukcji zdarzeń w CMS. Jest tu stosowany tzw. algorytm Particle Flow, który opiera się na połączeniu informacji z różnych podzespołów pozwalających na skuteczną rekonstrukcję cząstek w stanie końcowym. Autor opisuje dość szczegółowo jego działanie, prowadzące do uzyskania torów-kandydatów na elektrony i miony, klastrow w kalorymetrach, oraz późniejszego ich łączenia i eliminacji fałszywych torów. Efektem działania algorytmu jest lista cząstek i dżetów w stanie końcowym.

W dalszej części rozdziału czytelnik dowiadyuje się, jak rekonstruowane są semi-leptonowe (zwane na potrzeby pracy „hadronowymi”) rozpadły leptonu τ , a wreszcie jak rekonstruowana jest masa bozonu Higgsa, który rozpadł się w kanale $\tau\tau$, gdzie jeden z leptonów τ rozpada się następnie hadronowo, a drugi na mion i neutrino.

Rozdział ten jest napisany dość jasno, znajdują się tam jednak sformułowania, które wzbudziły moje wątpliwości:

str. 37: „First the PF charged hadron, PF neutral hadron or PF photon are extracted by comparing the (summed) track momentum and the calibrated calorimeter energy. At last, the PF neutral hadrons and PF photons are created using remaining clusters (not linked to any track).”

Do pędu jakiego toru porównywana jest energia zdeponowana w kalorymetrze w przypadku fotonów i hadronów neutralnych? Jak to się ma do drugiego zdania z cytatu?

str. 39: „The electrons and muons coming from the leptonically decaying taus are reconstructed by standard techniques [73, 74, 75, 76], and taken as they are.”

Nie jest tu jasne, co autor miał na myśli. Dokładniejsza informacja byłaby przydatna ze względu na to, że w dalszej analizie używane są właśnie przypadki z jednym z leptonów τ rozpadającym się na mion i neutrino.

Rozdział 4 przedstawia oficjalną analizę poszukiwania $H \rightarrow \tau\tau$, używaną w eksperymencie CMS i opartą na cięciach. Autor wykonał jej pełną powtórzną implementację, co było koniecznym krokiem do przeprowadzenia dalszych prac. Szczegółowo omówione zostały wykorzystane próbki danych i symulacji Monte Carlo oraz kolejne kroki w procesowaniu danych. Autor podaje szczegółowo wartości cięć czy wzory na obliczenie pewnych zmiennych używanych w selekcji, zabrakło natomiast ilustracji pokazujących efekty cięć czy rozkłady tych zmiennych. Załączenie takich wykresów pozwoliłoby łatwiej zrozumieć proces selekcji.

Rozdział ten zawiera również listę poprawek i normalizacji dokonanych na próbkach Monte Carlo, oraz krótki opis głównych źródeł tła. Omówione są również zmienne używane w końcowej selekcji, której dokonuje się na dwuwymiarowych rozkładach dla trzech kategorii przypadków. Na koniec pokazane jest porównanie wyniku implementacji dokonanej przez autora z oficjalnym wynikiem zespołu CMS, oraz krótkie zestawienie źródeł błędów systematycznych.

Rozdział 5 rozpoczyna się od informacji o istocie uczenia maszynowego i różnych jego rodzajach. Następnie autor przechodzi do opisu przygotowania próbek do procesu uczenia

wybranych klasyfikatorów (preselekcji), przeprowadzanej tak samo jak w rozdziale 4 (dla rozpadu $H \rightarrow \tau_\mu \tau_h$). Jednakże w nagłówku tabeli 5.1 prezentującej kategorie pojawia się zapis $H \rightarrow \tau_h \tau_h$ – co może być pomyłką, lub zmianą względem rozdziału 4, której opisu zabrakło.

W kolejnych podrozdziałach prezentowane są wybrane klasyfikatory: drzewo decyzyjne, losowy las decyzyjny i wzmacniane drzewo decyzyjne oraz sieć neuronowa. Przedstawione są ich zalety i wady, proces uczenia oraz metody oceny efektywności klasyfikacji zdarzeń. Rozdział ten jest dobrze i jasno napisany, a w zrozumieniu treści pomagają liczne rysunki i schematy.

Rozdział 6 prezentuje zastosowanie najlepszego klasyfikatora – sieci neuronowej do separacji sygnału $H \rightarrow \tau_\mu \tau_h$ od tła i porównanie tak otrzymanego wyniku z wynikiem analizy oficjalnej. Analizę przeprowadzono dla trzech kategorii przypadków, po cięciach opisanych w rozdziale 4, a jako kryterium porównawcze przyjęto znaczość wyniku (*significance*), zdefiniowaną w równaniu 6.2 i dodatkowo objaśnioną w dodatku E. Autor pokazał, że użycie sieci neuronowych zwiększa znaczość od 20 do 35% w zależności od kategorii.

Krótkie podsumowanie, poza przypomnieniem wyników, podaje też możliwe zastosowania w pomiarach konkretnych wartości fizycznych, oraz wspomina o przygotowywanej publikacji, w której metoda sieci neuronowych zostanie użyta. Zabrakło informacji, jak duży wkład do tej publikacji wniesie autor rozprawy.

Pięć dodatków dostarcza dodatkowych szczegółów, głównie technicznych, uzupełniających informacje zawarte w rozdziałach 4 i 6. W mojej opinii dodatki C (rozkłady zmiennych dla oficjalnej analizy) i E (objaśnienia dotyczące znaczości) mogłyby znaleźć się w głównej treści rozprawy.

Spis literatury obejmuje 168 pozycji, w większości stanowiących recenzowane publikacje. Można by dyskutować, czy nie ma lepszych źródeł do zacytowania przy opisie uczenia maszynowego niż artykuły z Wikipedii (pozycje 99, 111, 117 itd.) czy innych stron internetowych (100, 101, 116 itd.).

Praca napisana jest w dość trudny sposób. Opisy są bardzo szczegółowe, a nie wszystkie przytaczane szczegóły są istotne dla opisywanej analizy. Wyjątkiem są tu rozdziały 5 i 6, które są znacznie bardziej przejrzyste. Praca według mnie zyskałaby, gdyby – jak już wspomniałam – mgr Olszewski zamieścił w rozdziale 4 więcej wykresów pokazujących np. efekty opisywanych cięć, czy poświęcił więcej miejsca omawianym poprawkom poza zamieszczeniem licznych tabel. W orientacji wśród licznych skrótów pomaga na szczęście ich spis zamieszczony na początku pracy (choć nie wszystkie podane tam skróty znalazły się w treści rozprawy).

Autor posługuje się też czasem slangiem, przykładowo „hadronic tau lepton”, „visual mass”, czy określeniem „mass”, kiedy ma na myśli „invariant mass” (wyjaśnia to dopiero na str. 57).

Powyższe uwagi nie mają jednakże wpływu na pozytywną ocenę pracy, która przedstawia trudną i pracochłonną analizę, mogącą stanowić znaczące ulepszenie badań rozpadów $H \rightarrow \tau\tau$. Nie mam wątpliwości, że przedstawiona rozprawa w pełni spełnia warunki stawiane przez ustawę pracom doktorskim, dlatego też wnoszę o dopuszczenie mgr. Olszewskiego do dalszych etapów przewodu doktorskiego.

Z poważaniem
Justyna Łagoda

