Prof. dr hab. Piotr Salabura
Instytut Fizyki im. M. Smoluchowskiego
ul. Prof. Łojasiewicza 11
Uniwersytet Jagielloński
30-348 Kraków

8.12. 2022
Kraków

Recenzja pracy doktorskiej mgr inż. Macieja Majewskiego "Application of machine learning methods for the analysis of the calibration of the LHCb VELO detector, studies of irradiated silicon pixel sensors and reconstruction of the neutrino interactions in LARTPC detector "

Praca doktorska pana mgr inż. Macieja Majewskiego w swojej zasadniczej części dotyczy problematyki kalibracji detektora wierzchołka VELO (VErtex Locator) w eksperymencie LHCb pracującego na LHC w CERN. Główną misją eksperymentu LHCb jest poszukiwanie łamania symetrii CP w rozpadach hadronów z kwarkami b. Osiągnięcia eksperymentu obejmują także odkrycia nowych powabnych stanów (mezonów i barionów) oraz liczne wyniki w obszarze badań materii silnie oddziałującej. Cechą wyróżniającą spektrometr LHCb jest precyzyjny detektor wierzchołka, jeden z najlepszych na świecie , który w połączeniu z bardzo efektywnym system procesowania danych w czasie rzeczywistym umożliwia rekonstrukcje rozpadów poza punktem interakcji i filtrowanie interesujących rzadkich rozpadów przy olbrzymiej częstości interakcji (obecnie 40 MHz). Przedstawione w pracy opracowania dotyczą problemów kalibracji detektora VELO i monitorowania jej poprawności z zastosowaniem różnorodnych technik sztucznej inteligencji opartych na uczeniu maszynowym (machine learning).

Z uwagi na olbrzymią ilość kanałów kalibracja oraz monitorowanie pracy detektora krzemowego w LHCb (170 000 sensorów dla VeloStrip-wersja paskowa i ponad 40 mln dla VeloPix-wersja pixelowa) jest nietrywialnym zagadnieniem wymagającym odpowiedniego podejścia. Ponadto odróżnienie sygnału od szumów przy jednoczesnym utrzymaniu doskonałej wydajności ($\geq$ 99%) ma kluczowe znaczenie dla systemu wyzwalania detektora

opartego na synchronicznej rekonstrukcji śladów oraz dla możliwości archiwizacji danych. Prace nad rozwojem oprogramowania dla detektorów wierzchołka w LHCb są prowadzone od lat w grupie AGH kierowanej przez Prof. Tomasza Szumlaka. Pan mgr. inż. Maciej Majewski przedstawił w swojej pracy procedury oparte na uczeniu maszynowym umożliwiającej redukcje wymiarowości danych i efektywną analizę jakości kalibracji. Algorytm wykrywania sensorów o znacznie odbiegających progach dyskryminacji został włączony do procesu monitorowania pracy detektora w czasie pomiarów LHCb w 2018 r (tzw. Run2). W oparciu o wyniki tej analizy autor zaproponował algorytm, oparty na rekurencyjnych sieciach neuronowych, wykorzystujący trendy czasowe rozkładów szumów sensorów (ich szerokości) jako element wspomagający decyzję o konieczności ich zmian w czasie pomiarów. Podjęcie takiej decyzji o przeprowadzeniu nowego pomiaru kalibracyjnego ma istotne konsekwencje ponieważ wymaga zmiany systemy wyzwalania w LHCb (trygera) ora pracy detektora bez wiązki, co w warunkach pracy LHC nie jest trywialne. Zastosowanie proponowanej metody zostało przetestowane na próbce danych pomiarowych ale jeszcze nie wdrożone (być może będzie to miało miejsce w czasie Run3). Kolejnym istotnym osiągnieciem pracy jest opracowanie autorskiego algorytmu dla nowego detektora VeloPix wiążącego zależność długości trwania sygnału (tzw. czasu powyżej progu Time Over Threshold-TOT) z wielkością zaabsorbowanej fluencji promieniowania w sensorze w oparciu o funkcje zastępczą. Zaproponowana metoda może umożliwić określenie zaabsorbowanej fluencji w sensorach od mierzonej wartości TOT i pozwolić na oszacowanie zniszczeń radiacyjnych praktycznie w czasie rzeczywistym. Opracowana metoda została przetestowana na próbce danych uzyskanych przez naświetlanie prototypu detektora VeloPix.

Oprócz wymienionych wyżej analiz, związanych z detektorem VELO, autor pracy przedstawił projekt systemu Storck, opartego na interfejsie webowym oraz bazie danych, który umożliwia archiwizacje i pobieranie danych kalibracyjnych oraz system Titana do monitorowania stanu kalibracji. Oba projekty są złożonymi środowiskami programistycznymi opracowanymi przez zespół kierowany przez mgr. Inż. Macieja Majewskiego.

Ilość opracowanych algorytmów oraz biegłość w posługiwaniu się różnymi technikami uczenia maszynowego jest imponująca i zapewne wystarczyłaby na więcej niż jeden doktorat. Z drugiej strony, w moim odczuciu, część interpretacyjna uzyskanych wyników w kontekście fizyki detektora jest bardzo krótka i pozostawia pewien niedosyt którego efektem są uwagi zawarte w recenzji.

Praca ma ponad 150 stron, składa się z 8 rozdziałów i jest napisana w bardzo dobrym języku angielskim oraz listę publikacji (8) i wystąpień konferencyjnych (9) zawierających materiały związane z pracą . Praca zawiera pewne literówki lub braki w opisie rysunków, niezbyt liczne,

załączone do recenzji. Bardziej poważne niespójności lub braki w prezentacji wyników załączam poniżej w opisie poszczególnych rozdziałów.

Rozdział pierwszy zawiera zwięzłe wprowadzenie do modelu Standardowego oraz problematyki łamania symetrii CP. Rozdział 2 przedstawia strukturę spektrometru LHCb ze szczególnym uwzględnieniem detektora wierzchołka VELO w konfiguracji paskowej (VeloStriP), używanej poprzednio (w tzw. Run2), oraz zmodyfikowanej wersji pikselowej (VeloPix), zainstalowanej w tym roku dla pomiarów Run3, o znacznie większej segmentacji umożliwiającej pracę przy większej krotności interakcji. W rozdziale opisano procedurę kalibracji detektora oraz wykazano istotność monitorowania jej jakości w czasie pomiaru. Celem kalibracji jest wyznaczenia piedestałów oraz progów aktywujących przesył danych z każdego sensora, co ma zasadnicze znaczenie dla wielkości strumienia danych. Monitorowanie rozkładów progów ma także istotne znaczenie w celu określenia poziomu zniszczeń radiacyjnych, które mogą zostać częściowo skompensowane poprzez modyfikacje napięcia zasilania. Rozdział 3 przedstawia techniki uczenia maszynowego zastosowane w analizie danych. W szczególności pożyteczne były dla mnie paragrafy dotyczące: (i) użycia sieci neuronowych trybie nienadzorowanym (unspervised) , w trybie rekurencyjnym, redukcji wymiarowości problemu, metody wykrywania klastrów i ich klasyfikacji zastosowane w analizie detektora VELO (ii) metodu wzmocnionego uczenia maszynowego zastosowane w opracowaniu metody poszukiwania śladów w komorze projekcji czasowej LArTPC. Rozdział czwarty zawiera wprowadzenie do problemu kalibracji detektora VeloStrip z przykładami ilustrującymi rozkłady piedestałów oraz progów dyskryminacji szumów w sensorach typu R (współrzędne radialne) oraz Φ (współrzędne azymutalne) uzyskane z pomiarów kalibracyjnych. Dane zostały wykorzystane do stworzenia modelu opartego na uczeniu maszynowym (bez nadzoru) wykrywającego sensory z istotnie innym poziomem odcięcia (progu). Następnie, przedstawiono bardzo efektywne systemy redukcji wymiarowości, oparte na analizie głównych komponentów (Principal Component Analysis-PCA) i zastosowaniu auto-enkodera, umożliwiający skompresowana wizualizację trendów czasowych progów dyskryminacji (rys. 4.3.1-4.3.12). Zaletą przedstawionego podejścia jest łatwość obserwacji złej kalibracji sensorów oraz, przy pewnym doświadczeniu, śledzenie trendów zmian poszczególnych sensorów. Zaobserwowano brak wyraźnych trendów dla wartości średnich piedestałów ale bardzo wyraźne trendy dla ich szerokości (czyli rozkładów szumów). Te ostatnie zmieniają się z czasem (rys.4.1), czyli podążają za wzrostem poziomu zaakceptowanej fluencji. W oparciu o te obserwacje opracowano metodę rekurencyjnego uczenia maszynowego przewidującą nowe parametry kalibracji sensorów w oparciu o obserwowane trendy czasowe z pomiaru w 2018 roku (rys.4.4.3). Wyniki procedury wydają się być dobrze zbieżne z danymi, z wyjątkiem jednego okresu (ostatniego po prawej) czego powód

nie został przedyskutowany w pracy. W prezentacji wyników zabrakło dyskusji możliwych kryteriów decyzji o konieczności wykonania nowej kalibracji, która jak wskazałem powyżej ma zasadnicze znaczenie dla eksperymentu. Będzie to zapewne jednym z kluczowych aspektów dla kalibracji w nadchodzącym Run3. W kilku miejscach w pracy autor konkluduje iż zastosowane zmiany napięć na sensorach pozwoliły na zminimalizowanie efektów zniszczeń radiacyjnych choć nie zostało to w sposób jasny wykazane w pracy.

Rozdział 5 przedstawia opracowanie funkcji zastępczej wiążącej TOT z fluencja zaabsorbowaną przez sensory VeloPix i opracowaną na podstawie danych uzyskanych z prototypu detektora. Jak już pisałem na wstępie recenzji, taka funkcja może docelowo pozwolić na określenie poziomu zniszczeń radiacyjnych w sensorach praktycznie w czasie rzeczywistym. Funkcja jest wielomianem opisywanym przez 4 parametry, z których jeden (p1) charakteryzujący część liniową zależności (rys. 5.2,2) ma szczególnie eksponowaną zależność od fluencji (rysunek 5.2.7). Autor pracy przedstawił wiarygodny model statystyczny ewolucji parametrów od fluencji (zweryfikowany poprzez porównanie z danymi –rys.5.2.12-5.2.13). Model zawiera założenie o gaussowskim charakterze rozkładów parametrów, które jest dość dobrze spełnione z wyjątkiem rozkładów 4 parametru. Korelacje pomiędzy parametrami są eliminowane przez zastosowanie analizy PCA. Generowane w Monte Carlo rozkłady TOT (rys 5.2.15) oraz ewolucje zależności parametrów funkcji zastępczej od fluencji (rys.5.2.14) dobrze opisują obserwowane trendy. Dyskusja i prezentacja wyników dotyczących parametrów funkcji zastępczej nie są całkowicie spójne. Przyczyną wydaje się być zmiana numeracji binów fluencji (0-7) , zastosowana w tabeli 5.2.1 (i odpowiadający jej rys. 5.2.5), które jest dalszej części pracy odwrócona (rysunek 5.2.7). W konsekwencji kluczowy wynik pracy, zaprezentowany na rys.5.2.14, który prezentuje wzrost parametru p1 z fluencją mija się z konkluzją w pracy mówiącą o jego zmniejszaniu, co jest zgodne z oczekiwaniami (TOT powinno maleć z fluencją). Trochę pobocznym, aczkolwiek interesującym, aspektem przedstawionym w tym rozdziale (sekcja 5.1), jest badanie efektywności i poprawności rozpoznawania „szumiących" klastrów które są maskowane przy użyciu algorytmów OPTICS i DBscan. Wyniki skuteczności zaprezentowane w tabeli 5.1.1 (brak definicji macierzy dezorientacji –„confusion matrix") oraz przedstawiona dyskusja nie prowadzą do klarownej konkluzji co do tego które z podejść lepiej spełnia stawiane wymagania. Rozdział 5 przedstawia dwa środowiska służące celom kalibracji detektora VELO (Storck) oraz monitorowania jego działania(Titania). Stork jest systemem do archiwizacji danych kalibracyjnych i zawiera relacyjną bazę danych PostgresSQL (która także umożliwia nierelacyjny dostęp). Został napisanych przy użyciu języka python i webowego środowiska Django. Titania ma zastąpić używany dotychczas system monitorowania Lovel i jest napisana w języku python używając środowiska graficznego Qt. Przy pomocy utworzonego interfejsu

graficznego użytkownik uzyskuje dostęp do danych (zapisanych w różnych postaciach), może je przedstawiać w postaci wykresów oraz robić zestawianie rysunków pochodzących z różnych źródeł. W pracy przedstawiono architektury systemów, przykłady ich użycia ale w dość ogólnym przeznaczeniu, tzn. bez zastosowania do detektora VELO.

Dwa ostatnie rozdziały wychodzą poza główna tematykę pracy związaną z detektorem VELO i przedstawiają propozycję uniwersalnego podejścia do rekonstrukcji śladów powstających z interakcji neutrin rejestrowanych w komorach projekcji czasowej, wypełnionych ciekłym Argonem. Proponowany algorytm opiera się na zastosowaniu głębokiego, wzmocnionego uczenia maszynowego z zastosowaniem procesu Markowa z funkcjami nagród. Rozdział 9 przedstawia architekturę, formalizm matematyczny oraz jego implementację i przykłady wizualizacji procesu rekonstrukcji. System został napisany w środowiskach Python Package który zawiera środowisko uczenia maszynowego Openai-gy. System został przetestowany na podstawie danych symulacyjnych udostępnianych przez współpracę DeepLearnPhysics które zawierały oprócz sygnałów neutrinowych także procesy tłowe. Wykazano jego podstawową funkcjonalność oraz przedstawiono charakterystykę pracy sieci poprzez ewolucje funkcji strat oraz nagradzania. Natomiast, jak autor pisze, ewaluacja algorytmu pod kątem poprawności rekonstrukcji (czystości) i jego wydajności jest kwestia przyszłości i została powierzona przyszłym badaczom. Przedstawione opracowanie bardzo by zyskało gdyby zawarto w nim chociaż w minimalnym zakresie informacje i wydajności rekonstrukcji śladów, oraz ich typów, na podstawie danych symulacyjnych użytych do rozwoju systemu.

Podsumowując, stwierdzam iż przedstawiona do recenzji praca mgr. inż. Macieja Majewskiego w pełni spełnia warunki ustawy o stopniach naukowych i tytule naukowym doktora. Praca przedstawia wartościowe opracowania algorytmów oraz środowisk programistycznych dla monitorowania kalibracji detektora VELO dla eksperymentu LHCb. Dlatego też wnoszę do Rady Dyscypliny Nauk Fizycznych AGH o dopuszczenie mgr. inż. Macieja Majewskiego do dalszych etapów przewodu doktorskiego.

Prof. dr hab.

Piotr Salabura

Piotr Salabura

Lista sugerowanych poprawek edytorskich

Str.11 „as can be seen in the 1.3.2, the constraints" -as it can be seen in the 1.3.2, ..

Str 17: rysunek 2.3.2 nie ma oznaczenia, podobnie rysunek 2.4.1

Str 23: , .." H_t was set to be five times higher than the noise", 5 sigma above?

na tej samej stronie "low threshold", nie jest zdefiniowany

str. 27 ostatnie zdanie „..must be carefully study" , must be carefully studied

str.33 równanie 3.6 -symbol theta..

str. 76 – brak opisów osi X, Y na rys. 4.3.7- 4.3.10,

podobnie na rysunku 5.2.11

str. 7.3.11 symbol grecki β zamiast B

str. 132 ostatnie zdanie „"it changes the nucleon.." , it changes the charge of nucleon"

str.133- rysunek 7.3.4 – brak opisu (TODO) , oprocz tego jest problem z zachowaniem ładunku i liczby leptonowej (chyba zamiast mionu dodatniego powinien być elektron a mionu ujemnego pozyton o odpowiednie neutrina..)

str. 137 „"..from the traninig could be use for particle tracking ".. could be used.

Str. 141 ".is one of thre categories" .. three

Poniżej "The the last dimension" , jedno the za duzo

str. 144 .."which is eected" ??