Dr hab. inż. Zbigniew Świder, prof. PRz
Katedra Informatyki i Automatyki
Politechnika Rzeszowska im. Ignacego Łukasiewicza
al. Powstańców Warszawy 12
35-959 Rzeszów

Rzeszów, 26.07.2023

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy:	Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data
Autor rozprawy:	mgr inż. Wojciech Sikora
Promotor rozprawy:	prof. dr hab. inż. Joanna Polańska
Dziedzina:	nauki inżynieryjno-techniczne
Dyscyplina:	inżynieria biomedyczna

Niniejsza recenzja została przygotowana na zlecenie Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej.

1. Cel i zakres rozprawy

Rozprawa doktorska mgr inż. Wojciecha Sikory dotyczy szczegółowej analizy danych próbek pobranych od pacjentów z nowotworem głowy i szyi, w celu opracowania pełnego schematu dla przetwarzania danych obrazowania spektrometrii mas uzyskanych z próbek biologicznych, wykorzystującego elementy uczenia maszynowego na pozyskanym zestawie funkcji, a także zbadanie jakości przetwarzania danych. W wyniku przetwarzania dostajemy optymalny zestaw cech związanych z określonymi pikami w widmach masowych, które to mogą być wykorzystane do trenowania klasyfikatorów i odkrywania biomarkerów. Schemat ten można wykorzystać dla dowolnego zestawu danych pozyskanego za pomocą obrazowania spektrometrii mas MALDI-TOF, a uzyskany zestaw cech można wykorzystać do identyfikacji i kwantyfikacji białek oraz do przygotowania badania klinicznego.

W pracy postawiono trzy hipotezy. Pierwsza twierdzi, że identyfikacja pików w spektrach masowych może być skutecznie przeprowadzona za pomocą modelowania całego spektrum, poprzez podzielenie go na części oraz zamodelowaniu ich mieszaninami normalnymi. Druga hipoteza twierdzi, że informacja o przestrzennej dystrybucji danych pozyskana z obrazowania spektrometrią mass skutecznie usuwa redundancje i znacznie zmniejsza wymiarowość danych, przy jednoczesnym zachowaniu jakości danych. Ostatnia hipoteza twierdzi, że identyfikacja najważniejszych cech, dla danych heterogenicznych jest możliwa i skuteczna dzięki wnioskowaniu na podstawie wielu modeli jednostkowych.

2. Struktura i zawartość rozprawy

Recenzowana praca doktorska obejmuje formalnie sześć głównych rozdziałów poprzedzonych wstępem oraz zakończonych podsumowaniem. Zasadnicza część rozprawy liczy łącznie 105 stron, a ponadto zawiera spis bibliografii liczący 78 pozycji oraz spis rysunków i tabel.

Praca rozpoczyna się wstępem, w którym przedstawiono motywację, cele oraz tezy postawione w rozprawie. Zdaniem Autora, głównym celem rozprawy było przygotowanie solidnego schematu dla przetwarzania danych obrazowania spektrometrii mas uzyskanych z próbek biologicznych pobranych od pacjentów z rakiem. Zastosowanie uczenia maszynowego na pozyskanym zestawie funkcji pozwoliło na wytrenowanie dobrze działających klasyfikatorów oraz na ulepszenie procesu modelowania widma na etapie detekcji pików. Proponowany schemat zapewnia również zestaw cech, które można wykorzystać do identyfikacji i kwantyfikacji białek, a następnie do przygotowania badania klinicznego.

W rozdziale 2 przypomniano podstawowe informacje na temat samej spektrometrii mas oraz obrazowania z jej wykorzystaniem. Pokrótce wyjaśniono znaczenie analizy danych obrazowych próbek biologicznych oraz wprowadzono główne pojęcia z tym związane. Wyjaśniono też proces pozyskiwania widm masowych i różnice między technikami spektrometrii mas.

Rozdział 3 zawiera szczegółowy opis danych wykorzystanych podczas eksperymentów. Opisano krok po kroku proces przygotowania próbek oraz akwizycji danych, wraz ze wszystkimi ważnymi parametrami i specyfikacją zastosowanego spektrometru masowego. Przedstawiono również wstępne kroki niezbędne dla przygotowania surowych danych na potrzeby identyfikacji pików.

W rozdziale 4 zawarto przegląd obecnie stosowanych metod detekcji pików dla danych z wykorzystaniem spektrometrii mas. Ich celem jest identyfikacja wszystkich pików w sygnale, które odpowiadają rzeczywistemu składnikowi, a także i przekształcenie dużej grupy punktów danych w odpowiednio mały zestaw cech, wystarczający dla dalszej analizy. Po wykryciu piku dane te mogą być dalej analizowane, np. w celu odkrycia lub klasyfikacji biomarkerów.

Rozdział 5 zawiera opis wykorzystanej metody wykrywania pików za pomocą modelowania widma w oparciu o mieszaniny Gaussa. Kolejno opisano stosowane metody podziału widma na mniejsze fragmenty, a następnie proces dopasowywania modeli mieszanin Gaussa do fragmentów wraz ze szczegółowym opisem algorytmu oraz procesem wyboru optymalnej liczby składników mieszanki. Na końcu rozdziału przedstawiono wyniki implementacji metody detekcji pików do danych testowych.

Rozdział 6 dotyczy inżynierii cech, której celem jest redukcja wymiarowości od zestawu tysięcy Gaussowskich składowych modelu widma do małego zestawu cech wykorzystujących rzeczywistą wiedzę o procesie akwizycji danych MSI. Zastosowano tu autorskie wykorzystanie przestrzennego rozkładu cech w celu ułatwienia całego procesu inżynierii cech. W rozdziale podano szczegółowy opis zastosowanych metod oraz całego procesu inżynierii funkcji.

W rozdziale 7 przedstawiono zastosowanie metod statystycznych i uczenia maszynowego do szkolenia klasyfikatorów zdolnych do przewidywania nowej obserwacji na podstawie przetworzonych danych. Opisano strategię podziału zbiorów danych na zestawy szkoleniowe, testowe i walidacyjne, a także obszerny zestaw narzędzi służący do porównywania modeli i oceny wydajności.

W rozdziale 8 opisano algorytmy używane do szkolenia klasyfikatorów, a w szczególności oparty na regresji logistycznej oraz wykorzystujący sieci neuronowe. Następnie porównano wydajności klasyfikacji dla obu tych algorytmów oraz przedstawiono szczegółowy opis oceny ważności cech.

W podsumowaniu omówiono wyniki przeprowadzonych eksperymentów, wyciągnięte wnioski oraz plany przyszłych badań.

3. Najważniejsze osiągnięcia rozprawy

Biorąc pod uwagę zawartość pracy oraz pozytywna ocenę jej treści merytorycznej, za główne osiągnięcia Autora należy uznać opracowanie szczegółowego schematu przetwarzania danych obrazowania spektrometrii mas uzyskanych z próbek biologicznych pobranych od pacjentów z rakiem oraz wykonanie szczegółowych badań wraz z prezentacją uzyskanych wyników. Należy zauważyć, że Autor podjął się realizacji bardzo ciekawego oraz istotnego z punktu widzenia praktycznych zastosowań tematu badawczego.

Najważniejszymi elementami rozprawy, decydującymi o jej wartości naukowej i badawczej, są:

1. Opracowanie procesu analizy danych MALDI-TOF MSI ze szczegółowymi etapami wykrywania pików, filtrowania szumów oraz metod inżynierii cech, które ostatecznie prowadzą do dobrze zdefiniowanego i nieredundantnego zestawu funkcji, które można wykorzystać do szkolenia poprawnie działających klasyfikatorów.
2. Zbadanie procesu podziału widma dla modelowania jego poszczególnych części oraz propozycja wskaźników określających cele tego procesu i propozycja nowej metody podziału tegoż widma.
3. Nowatorskie zastosowanie podejmowania decyzji w oparciu o rozkład przestrzenny podczas inżynierii cech danych MSI oraz podczas procesu wykrywania otoczki izotopowej.

4. Poprawność pracy i uwagi krytyczne

Poprawność treści rozprawy nie wzbudza zastrzeżeń, a stwierdzenia w niej zawarte stanowią podstawę do kontynuowania badań, co wynika w szczególności z przedstawionych podstaw teoretycznych popartych wynikami przeprowadzonych badań eksperymentalnych.

Jednocześnie Autor nie ustrzegł się pewnych drobnych niedociągnięć, a wśród uwag o charakterze krytycznym, a po trosze i dyskusyjnym, można wymienić następujące:

1. Czy proponowany schemat postępowania daje znaczną przewagę (jeśli tak, to w czym) nad innymi algorytmami opisywanymi w literaturze? Czy opisana w pracy modyfikacja algorytmu EM została przetestowana na innych danych (w celu potwierdzenia jej skuteczności)?
2. Zdaniem Autora proponowany proces redukcji z wykorzystaniem metod statystycznych prowadzi do znacznego zmniejszenia wymiarowości danych. A czy to samo można uzyskać innymi metodami? Jaka jest zaleta takiego (proponowanego) podejścia?
3. Czy ocena ważności cech dla zastosowanych klasyfikatorów jest zbieżna z ogólną oceną specjalisty (lekarza analizującego te zdjęcia)? Wyniki testów przedstawione pracy różnią się zwykle o kilka punktów procentowych, więc raczej nie można twierdzić, że któraś metoda jest wyraźnie lepsza od pozostałych (zależy to też w dużej mierze od danych wejściowych).
4. Czy proponowane rozwiązanie zostało zaimplementowane w formie gotowej aplikacji, możliwej do zastosowania w badaniach medycznych? Czy taka aplikacja może być w pełni zautomatyzowana (uruchamiamy na dowolnym zbiorze danych bez ingerencji lekarza), czy też trzeba za każdym razem dobierać pewne parametry wejściowe kierując się wiedzą ekspercką, aby otrzymać poprawne wyniki?
5. Drobne uwagi szczegółowe (najważniejsze):
   • Str. 5, 17, 21, 25, … – niejednakowa (różnej wielkości) czcionka użyta na rysunkach.
   • Str. 17, 24, 62, 68, … – trochę zbyt mała czcionka w opisach na rysunkach.
   • Str. 48 – niepotrzebnie pogrubiona cała linia (a nie tylko słowo kluczowe) w pseudokodzie.
   • Str. 50 – brak przecinka (na końcu) we wzorze (11).
   • Str. 79, 81, 83, … – tabela wystaje poza marginesy tekstu (dotyczy wszystkich dużych tabel).

5. Podsumowanie

Przytoczone wyżej uwagi dyskusyjne nie umniejszają zasług Autora ani nie kwestionują przedstawionych osiągnięć, a opisywana w pracy problematyka dotyczy aktualnych i interesujących zagadnień naukowych. Recenzowana praca zasługuje na pozytywną ocenę merytoryczną i wnosi istotny oraz oryginalny wkład w dyscyplinę inżynieria biomedyczna. Postawione cele i zadania pracy zostały zrealizowane, a jej tematyka wpisuje się we współczesny nurt badań w tym zakresie.

Stwierdzam zatem z pełnym przekonaniem, że opiniowana rozprawa mgr inż. Wojciecha Sikory pt. „*Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data*" zawiera samodzielne rozwiązanie ważnego i istotnego problemu naukowego. Recenzowana praca doktorska spełnia wymagania stawiane rozprawom doktorskim, zgodnie z Ustawą z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, art. 187 (Dz. U. z 2023 r. poz. 742, 1088, 1234), w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie inżynieria biomedyczna, tak więc wnoszę o przyjęcie rozprawy i jej dopuszczenie do publicznej obrony.

Dr hab. inż. Zbigniew Świder