Prof. dr hab. Zbisław Tabor
Katedra Biocybernetyki i Inżynierii Biomedycznej
Wydział Elektrotechniki, Automatyki, Informatyki
i Inżynierii Biomedycznej
Akademia Górniczo-Hutnicza
ztabor@agh.edu.pl

Kraków, 6 marca 2023

RECENZJA

rozprawy doktorskiej Pani mgr Anny Marcisz pt.:

„MODELE PREDYKCYJNE WSPOMAGAJĄCE DIAGNOSTYKĘ CHOROBY ALZHEIMERA ORAZ ŁAGODNYCH ZABURZEŃ POZNAWCZYCH Z WYKORZYSTANIEM OBRAZOWANIA METODĄ REZONANSU MAGNETYCZNEGO"

wykonanej pod kierunkiem Pani promotor prof. dr hab. Joanny Polańskiej

I. Ogólna charakterystyka podjętych przez Doktorantkę problemów badawczych

W przedłożonej mi do oceny rozprawie Doktorantka przedstawia założenia badawcze, metodykę oraz wyniki prac nad problemem badawczym, którego cel stanowi zaprojektowanie, implementacja i weryfikacja narzędzia służącego do wspomagania badań przesiewowych w diagnostyce choroby Alzheimera oraz łagodnych zaburzeń poznawczych. Planowane narzędzie ma, wykorzystując łatwo mierzalne cechy charakteryzujące diagnozowanego pacjenta, zaklasyfikować go do jednej z trzech kategorii: pacjent w normie, pacjent z łagodnymi zaburzeniami poznawczymi, pacjent z chorobą Alzheimera. Podjęty problem badawczy wpisuje się w obszar zastosowania metod informatycznych we wspomaganiu diagnostyki medycznej, a tym samym w obszar dyscypliny „inżynieria biomedyczna".

Podjęty problem badawczy jest rozwiązywany z wykorzystaniem metod analizy obrazu, uczenia maszynowego oraz statystyki. Doktorantka, projektując swoje rozwiązanie, wykorzystuje otwarte bazy danych, zawierające badania obrazowe (rezonans magnetyczny głowy), dane kliniczne (m.in. wyniki testu MMSE) oraz diagnozę, użytą jako dane referencyjne w treningu modeli klasyfikacyjnych.

Podjęty problem badawczy jest ważny szczególnie w kontekście najbardziej aktualnych problemów zdrowotnych starzejących się społeczeństw krajów wysoko

rozwiniętych. Znalezienie wiarygodnego i klinicznie użytecznego rozwiązania tego problemu stanowi niewątpliwie zadanie dla inżynierii biomedycznej.

II. Ogólna charakterystyka rozprawy

Na przedłożoną mi do oceny rozprawę składa się, oprócz wstępu i końcowych, podsumowujących części i spisów, pięć rozdziałów, których kolejność i zawartość odpowiadają schematowi pracy naukowej.

W rozdziale drugim omówiona jest ogólna charakterystyka podjętego przez Doktorantkę problemu badawczego: Doktorantka opisuje, czym jest choroba Alzheimera, jej objawy, przebieg, rokowania, terapię. Omówione są również metody diagnozowania, w tym metody wykorzystujące diagnostykę obrazową. Doktorantka omawia również skrótowo metody uczenia maszynowego używane w rozwiązywaniu zagadnień klasyfikacji. Omówienie to jest skrótowe z konieczności – szczegółowe omówienie tematyki klasyfikatorów wymagałoby z pewnością tekstu o objętości równej co najmniej objętości rozprawy, przedstawionej mi do oceny.

Rozdział trzeci i czwarty zawierają opis danych i metodyki użytych w zaproponowanym przez Doktorantkę rozwiązaniu problemu klasyfikacji trójklasowej. Dane pochodzą z dwóch otwartych baz danych, udostępniających badania obrazowe (rezonans magnetyczny) mózgu oraz dane uzyskane w wyniku wywiadu (wiek, płeć, wynik testu MMSE, wykształcenie). Metodyka użyta w pracy obejmuje analizę ilościową badania RM mózgu w oparciu o wcześniejszą segmentację mózgu oraz płynu mózgowo rdzeniowego. Jako narzędzie do rozwiązania problemu klasyfikacji trójklasowej Doktorantka wybrała wielomianową regresję logistyczną.

Rozdział piaty przedstawia wyniki uzyskane dla analizowanych zbiorów danych, a w rozdziale szóstym wyniki uzyskane przez Doktorantkę są porównane z opublikowanymi wynikami, uzyskanymi dla tego samego problemu lub dla problemów zbliżonych choć niekoniecznie dla tych samych zbiorów danych.

Cel oraz teza są sformułowane w rozdziale pierwszym rozprawy. Ponieważ cel zacytowałem już w rozdziale pierwszym tej recenzji, w tym miejscu przytoczę tylko tezę rozprawy: „Model predykcyjny wykorzystujący wyniki MMSE, wiek oraz względną objętość mózgu jest skutecznym testem przesiewowym, który efektywnie i z większym wyprzedzeniem czasowym, w porównaniu do testu opartego o MMSE, rozpoznaje chorych z łagodnymi zaburzeniami funkcji poznawczych, którzy później rozwiną pełnoobjawową chorobę Alzheimera."

Układ rozprawy uważam za prawidłowy, a bibliografię za wystarczającą.

III. Formalna ocena rozprawy

Jednym z moich zadań, jako recenzenta jest wyrażenie opinii, czy „rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie albo dyscyplinach oraz umiejętność samodzielnego prowadzenia pracy naukowej" (Art. 187 ust. 1 ustawy Prawo o szkolnictwie wyższym i nauce). Dostarczona mi dokumentacja (rozprawa i lista publikacji Doktorantki) przekonuje mnie, że Doktorantka zaproponowała koncepcję badań opisanych w rozprawie oraz była w te badania zaangażowana we wszystkich etapach – od etapu skompletowania danych, poprzez opracowanie metodyki, obejmującej rozwój narzędzi do analizy danych, w tym narzędzi opartych o uczenie maszynowe, do etapu przygotowania tekstu opisującego wyniki przeprowadzonych badań. **W oparciu o przekazaną mi dokumentację wyrażam zatem przekonanie, że oceniana przeze mnie rozprawa doktorska prezentuje ogólną wiedzę teoretyczną Doktorantki w dyscyplinie inżynieria biomedyczna oraz umiejętność samodzielnego prowadzenia pracy naukowej.**

Problematyka dotycząca komputerowego wspomagania diagnostyki choroby Alzheimera jest rozwijana od szeregu lat. Doktorantka mimo wszystko podejmuje ten ważny temat i projektuje oryginalne, efektywne kosztowo rozwiązanie w oparciu o – co istotne – otwarte zbiory danych, co pozwala w toku ewentualnych przyszłych badań na dokonanie wiarygodnych porównań uzyskanego rozwiązania z innymi, opublikowanymi rozwiązaniami tego samego problemu. **Rozprawa zawiera – moim zdaniem - wymagane przez ustawę Prawo o szkolnictwie wyższym i nauce (Art. 187 ust. 2) dla pozytywnej oceny rozprawy doktorskiej oryginalne rozwiązanie podjętego przez Doktorantkę problemu naukowego z obszaru dyscypliny inżynieria biomedyczna.**

IV. Uwagi do rozprawy

Projektując swoje narzędzie Pani mgr Anna Marcisz nie uniknęła nieścisłości metodycznych, z których najważniejsze, moim zdaniem, wymieniam poniżej.

1. Na stronie 38 jest napisane: „Otrzymane rezultaty były bardzo dobre i ostatecznie to narzędzie wybrano do ekstrakcji mózgu". Jakość procedur segmentacji mierzy się w analizie obrazu ilościowo. Rozumiem, że bazy danych

z których korzystała Doktorantka nie zawierają referencyjnych segmentacji mózgu, ale uzyskanie takich segmentacji, przynajmniej dla ograniczonego zbioru badań (np. 30 pacjentów), w celu ilościowej oceny metody segmentacji z pewnością nie jest zadaniem niewykonalnym.

2. Na str. 40 przywołany jest Gaussian Mixture Model (GMM) w celu opisu histogramu sygnału obrazu rezonansu magnetycznego (RM). Sygnał z RM nie podlega jednak rozkładowi Gaussa (odpowiednim rozkładem dla RM jest „Rician distribution").

3. Doktorantka przedstawia dość złożony algorytm w celu segmentacji płynu mózgowo-rdzeniowego. W wyniku wykonania algorytmu wyznaczany jest globalny próg segmentacji. Segmentacja płynu mózgowo-rdzeniowego jest więc w istocie rzeczy binaryzacją opartą o progowanie z globalnym progiem. Nie jest jasne, czy inne algorytmy służące do wyznaczania globalnych progów segmentacji (np. Otsu) nie zachowałyby się równie dobrze, jak algorytm użyty przez Doktorantkę. Jakość segmentacji płynu mózgowo-rdzeniowego nie jest w żaden sposób skwantyfikowana.

4. Na str. 52 Doktorantka raportuje wyniki testu normalności rozkładu wyników pomiarów. Nie jest napisane, jaki test statystyczny został użyty. Być może wystarczyłoby zaprezentować wykres kwantylowy, zamiast wykonywać formalny test normalności. O przeprowadzeniu testu równości wariancji (jedno z założeń testu ANOVA) Doktorantka już nie wspomina – można o tej równości wnioskować jedynie z wykresów pudełkowych i z tabel. Test ANOVA jest użyty przez Doktorantkę do analizy danych z badania obserwacyjnego. Oznacza to, że widzimy co prawda w analizowanych danych koincydencję między zdiagnozowaną chorobą Alzheimera, a zmniejszoną względną objętością mózgu, ale nie możemy tylko na tej podstawie wnioskować o związku przyczynowo-skutkowym (to byłoby możliwe w eksperymencie). Te same uwagi dotyczą wyników przedstawionych w rozdziałach 4.1.2.2. i 4.1.3.2.

5. W rozdziale 4.1.2. Doktorantka przedstawia algorytm wyznaczania „powierzchni mózgu bez płynu mózgowo-rdzeniowego". Kontury wyznaczane przez Doktorantkę autorskim algorytmem można alternatywnie wyznaczyć korzystając z gotowych funkcji bibliotek scikit-image lub openCV (google: „openCV find contour", „scikit-image find contour"). Usunięcie obiektów, nazwanych przez Doktorantkę „czarnymi dziurami" i „białymi plamami" można również alternatywnie zrealizować, korzystając z ww. bibliotek, rozpoczynając od etykietowania składowych spójnych, a następnie ich filtrowania według objętości

(google: „scikit-image label", objętości można policzyć funkcją unique z pakietu numpy języka python). Powierzchnię wysegmentowanej bryły można wyliczyć od razu w 3D, korzystając z algorytmu „marching cubes" (wyznaczanie powierzchni bryły 3D realizują dwie funkcje scikit-image, wywołane w sekwencji: marching_cubes i mesh_surface_area).

6. Jakość segmentacji komór bocznych (rozdział 4.1.3.1.) nie jest skwantyfikowana.

7. Do rozwiązania problemu klasyfikacji Doktorantka wykorzystuje modele wielomianowej regresji logistycznej, które pozwalają oszacować prawdopodobieństwo przynależności obiektu do jednej z klas na podstawie wartości funkcji liniowej pewnego zbioru zmiennych niezależnych, charakteryzujących diagnozowanego pacjenta. Ze zbioru wszystkich zmiennych niezależnych wyznaczanych dla pacjenta (wiek, płeć, liczba lat edukacji, punkty MMSE, względna objętość mózgu, współczynnik pofałdowania, objętość komór bocznych) selekcjonowany jest podzbiór, w oparciu o który budowany jest najlepszy (w jakimś sensie) model predykcyjny. Do wyboru najlepszego modelu Doktorantka wykorzystuje Bayesowskie kryterium informacyjne (BIC). Jest to dość ekonomiczne podejście, jeśli chodzi o wykorzystanie zasobów obliczeniowych. Z drugiej strony Doktorantka trenuje nie jeden, ale pięć modeli, korzystając z metodyki walidacji krzyżowej. Ale walidacja krzyżowa jest nowocześniejszą (i bardziej kosztowną obliczeniowo) alternatywą do BIC w zadaniu selekcji modelu. Skoro więc walidacja krzyżowa została i tak już wykonana, to o BIC można z czystym sumieniem zapomnieć i wykorzystać wyniki walidacji krzyżowej do selekcji modelu (tak, jak to się robi np. korzystając z funkcji GridSearchCV ze scikit-learn).

8. Do rozwiązania problemu klasyfikacji Doktorantka wykorzystuje modele wielomianowej regresji logistycznej – nie jest jasne, dlaczego całe bogactwo innych narzędzi do klasyfikacji zostało przez Doktorantkę pominięte. Być może chodzi o łatwość interpretacji wyników regresji logistycznej (ilorazy szans)? Niemniej, byłoby ciekawe dowiedzieć się, czy inne klasyfikatory poradziłyby sobie z problemem równie dobrze.

9. Doktorantka wykorzystuje do oceny modeli klasyfikacji bardzo wiele różnych wskaźników. Moim zdaniem wystarczyłyby krzywe ROC, ponieważ pozostałe wskaźniki zależą od wyboru progu decyzyjnego tj. od procedury, której jednoznacznego sprecyzowania w rozprawie nie zauważyłem.

10. Na str. 68 Doktorantka pisze w ostatnim zdaniu rozdziału 4.2: „Porównano model podstawowy z modelem rozszerzonym za pomocą ANOVA". Moim zdaniem jest to dość ryzykowne podejście, zważywszy, że wyjściem z obu modeli jest zmienna kategoryczna (0 lub 1), co stoi w jawnej sprzeczności z założeniem analizy ANOVA o zmiennej zależnej, która dla dla każdego poziomu czynnika podlegać ma rozkładowi normalnemu. Celem doktorantki jest porównanie dwóch modeli klasyfikacyjnych: najprościej jest wykonać takie porównanie, stosując test statystyczny porównujący pola pod krzywymi ROC dla dwóch porównywanych klasyfikatorów (np. http://vassarstats.net/roc_comp.html, google: „compare two roc curves").

11. Na str. 69 Doktorantka pisze: „BIC=785.54 dla modelu bez dodatkowych zmiennych versus BIC=809.68 dla modelu z dodanymi zmiennymi". Zwykle, korzystając z BIC, spośród kilku modeli wybiera się jako najlepszy ten, który ma niższą wartość BIC...

12. Tabela 13 na str. 77 prezentuje wyniki dla modelu rozszerzonego, wśród których są pokazane wartości AUC ROC, punkt odcięcia oraz inne wskaźniki jakości klasyfikacji. Nie jest jasne, jak został wybrany punkt odcięcia. Powinien być wybrany (w oparciu o konkretne kryterium np. wartość statystyki Youdena) dla każdego foldu walidacji pięciokrotnej dla zbioru treningowego i następnie zaaplikowany do zbioru walidacyjnego. Czułość, swoistość itd. dla zbioru walidacyjnego powinny być wyznaczone dla tak wybranego progu decyzyjnego.

13. W tekście brakuje tabeli analogicznej do tabeli 13, ale z wynikami dla modelu bazowego. Są tylko pokazane (w tabeli 10) macierze błędów, ale nie jest jasne dla jakiego progu odcięcia i jak wybranego są to wyniki. Nie ma więc możliwości porównania dla zbioru ADNI na podstawie obiektywnych przesłanek (AUC ROC) jakości klasyfikacji z wykorzystaniem modelu bazowego i rozszerzonego. Takie porównanie AUC ROC dla modelu podstawowego i rozszerzonego przedstawiono w Tabeli 16 dla zewnętrznego zbioru testowego EDSD – z porównania wynikałoby, że jakość klasyfikacji jest dla obu modeli, statystycznie rzecz ujmując, taka sama. Jedyna większa (4%) różnica AUC ROC między modelami jest obserwowana przy wartości AUC ROC ok. 70%, co najprawdopodobniej (należałoby przeprowadzić test statystyczny) nie jest statystycznie istotne.

14. Na stronach 85-89 Doktorantka przedstawia dodatkowe argumenty, które miałyby przemawiać na rzecz skuteczności modelu. Jeśli dobrze zrozumiałem tą argumentację, można by ją streścić tak: „Co prawda model predykcyjny, wytrenowany do postawienia diagnozy pacjenta, aktualnej na chwile obecną,

błędnie sklasyfikował pacjenta, ale stan pacjenta uległ po jakimś (dłuższym) czasie zmianie, więc ta błędna diagnoza na chwilę obecną okazała się poprawna po dłuższym czasie". Model predykcyjny został wytrenowany do stawiania diagnozy na chwile obecną, nie ma – moim zdaniem – żadnego uzasadnienia do używania tego modelu w innych kontekstach. Model przedstawiony przez Doktorantkę na str. 90 jest jedną z możliwości wnioskowania o przyszłym stanie pacjenta. Analogicznie, można by wytrenować model, który w oparciu o cechy wyznaczone w chwili obecnej przewiduje przyszły stan pacjenta (nie zmianę statusu, jak to zrobiła Doktorantka, ale właśnie status). W przypadku modelu przedstawionego na str. 90 chętnie zobaczyłbym krzywe ROC dla modelu bazowego i rozszerzonego. Chętnie dowiedziałbym się również, jak został wybrany zestaw cech, użytych w modelu rozszerzonym.

15. Dyskusja zawiera obszerne porównanie uzyskanych wyników z wynikami prac już opublikowanych. Wszelkie wnioski o przewadze jednego lub drugiego modelu nie mogą być uprawnione, dopóki modele nie zostaną wytrenowane na tym samym zbiorze danych, w tym samym schemacie poszukiwania hiperparametrów, selekcji progów decyzyjnych itd.

V. Wnioski

W mojej opinii przedłożona mi do recenzji rozprawa spełnia wszystkie, określone ustawą Prawo o szkolnictwie wyższym kryteria, wymagane do jej pozytywnej oceny:

1. prezentuje ogólną wiedzę teoretyczną Doktorantki w dyscyplinie inżynieria biomedyczna oraz umiejętność samodzielnego prowadzenia pracy naukowej,
2. zawiera oryginalne rozwiązanie podjętego przez Doktorantkę problemu naukowego, dotyczącego opracowania narzędzia wspomagającego diagnostykę choroby Alzheimera.

Oceniając zatem rozprawę pozytywnie, wnioskuję o dopuszczenie Pani mgr Anny Marcisz do dalszych etapów postępowania w sprawie nadania stopnia doktora w dyscyplinie inżynieria biomedyczna.

Signed by / Podpisano przez:
Zbisław Tabor
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Date / Data: 2023-03-06 19:27

Zbisław Tabor

Zbisław Tabor