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, 10 maja 2023

RECENZJA

rozprawy doktorskiej Pani mgr Katarzyny Sieradzkiej pt.:

„CLASSIFICATION OF WHITE BLOOD CELLS BASED ON SINGLE-CELL SEQUENCING DATA FOR BIODOSYMETRY PURPOSES”

wykonanej pod kierunkiem Pani promotor prof. dr hab. Joanny Polańskiej
i promotora pomocniczego Pana dra Christophe Badie

I. Ogólna charakterystyka podjętych przez Doktorantkę problemów badawczych

W przedłożonej mi do oceny rozprawie Doktorantka przedstawia założenia, metodykę oraz wyniki prac nad zagadnieniem badawczym, którego celem jest opracowanie metody wykorzystania wyników sekwencjonowania RNA pojedynczych komórek (białych ciałek krwi - WBC) w dwóch zadaniach:

1. rozróżnienie typów WBC,
2. rozróżnienie WBC napromienionych promieniowaniem jonizującym od WBC, które nie były napromienione.

WBC zostały wyizolowane z dwóch próbek krwi. WBC wyizolowane z każdej próbki zostały podzielone na grupę kontrolną i grupę poddaną działaniu promieniowania. Materiałem badawczym są więc dwie grupy kontrolne i dwie grupy poddane działaniu czynnika. Projektując rozwiązania powyższych zadań Doktorantka pracuje z surowymi danymi sekwencjonowania, które dla każdej z czterech grup WBC mają formę dwuwymiarowej macierzy zliczeń. W macierzy tej wiersze numerują pojedyncze sekwencjonowane komórki, a kolumny numerują poszczególne geny – komórka macierzy zawiera liczbę genów danego typu (skojarzonego z indeksem kolumny) znalezionych w procesie sekwencjonowania w danym WBC (skojarzonym z indeksem wiersza). Doktorantka operuje w swojej

pracy na czterech takich macierzach, każda licząca ok. 2000 do 3000 wierszy i ok. 400 kolumn.

Problem rozróżnienia WBC napromienionych promieniowaniem jonizującym od WBC, które nie były napromienione jest rozwiązywany przy użyciu metod uczenia maszynowego i statystyki, poprzedzonych metodami wstępnego przetwarzania surowych danych sekwencjonowania w celu zredukowania liczby zmiennych, a w zasadzie w celu znalezienia zmiennych istotnych dla rozwiązania problemu klasyfikacji. Znalezienie klasyfikatora charakteryzującego się wysoką dokładnością jest jednocześnie rozwiązaniem zadania znacznie bardziej istotnego z punktu zastosowań radiobiologicznych tzn. zadania wskazania tych genów, które pozwalają odróżnić komórki napromienione od komórek z grupy kontrolnej. Wskazanie takich genów może mieć istotne znaczenie dla poznania mechanizmów związanych z reakcją WBC na promieniowanie jonizujące, co z kolei może mieć przełożenie na ochronę radiologiczną, terapię wykorzystującą promieniowanie jonizujące itd.

WBC nie stanowią homogenicznej grupy komórek krwi, ale dzielą się na wiele różnych typów. Mechanizmy reakcji różnych typów WBC na promieniowanie jonizujące mogą być różne wobec czego typ WBC jest niekontrolowaną zmienną zaburzającą predykcję klasyfikatora, który ma za zadanie odróżnić komórki napromienione od kontrolnych. Aby usunąć wpływ typu WBC na model klasyfikacyjny, Doktorantka, korzystając z metod uczenia nienadzorowanego, dzieli wiersze macierzy zliczeń na klastry, które następnie – w oparciu o przesłanki biochemiczne (profil genetyczny) – kojarzy z typami WBC. To skojarzenie z konieczności (brak etykiet niezbędnych do przeprowadzenia treningu nadzorowanego) jest obarczone pewną dozą niepewności. Po wykonaniu klasteryzacji danych w macierzach zliczeń Doktorantka ponownie trenuje klasyfikator, tym razem tylko dla największego klastra danych, uzyskując wyniki, które są konsystentne z wynikami uzyskanymi poprzednio dla całego zbioru danych.

W swoich badaniach Doktorantka wykorzystuje regresję logistyczną, jako metodę klasyfikacji, co jest wyborem racjonalnym zważywszy, że zbiór danych treningowych nie jest jednak zbyt liczny. Jako alternatywną metodę klasyfikacji Doktorantka przetestowała sieci neuronowe, uzyskując wyniki zbliżone do tych uzyskanych przy użyciu regresji logistycznej.

Podjęty problem badawczy wpisuje się w obszar zastosowania metod informatycznych w radiobiologii, a tym samym w obszar dyscypliny „inżynieria biomedyczna”.

II. Ogólna charakterystyka rozprawy

Przedłożona mi do oceny rozprawę ma układ publikacji naukowej: po wstępie następuje omówienie materiału badawczego i metod badawczych, prezentacja uzyskanych wyników w zadaniach selekcji zmiennych objaśniających, klasyfikacji i klasteryzacji, rozprawę kończy dyskusja. Bibliografia liczy 108 pozycji. Do rozprawy dołączone jest sporo materiałów uzupełniających w formie wykresów i tabel, które mają uzasadniać dokonane przez Doktorantkę skojarzenie klastrów danych w macierzach zliczeń z typami WBC.

W trójstronicowym wstępie, sformułowane zostały cele pracy, które wymieniłem w pierwszym rozdziale tej recenzji. We wstępie sformułowane zostały również tezy, które cytuję w całości:

1. Combining feature engineering methods and advanced dimensionality reduction techniques with unsupervised clustering algorithms allows for the efficient identification of white blood cell subtypes in single-cell RNA sequencing data.

2.The proposed intelligent and stratified algorithm of the training set construction supports the classification system, especially in the case of heterogeneous dataset.

Układ rozprawy uważam za prawidłowy, a bibliografię za wystarczającą.

III. Uwagi do rozprawy

Tą część recenzji rozpocznę od uwagi drobnej, językowej, ale mającej zastosowanie do zbyt wielu fragmentów tekstu rozprawy. Doktorantka ma, moim zdaniem, skłonność do nadużywania przymiotników, co niekoniecznie jest właściwe w tekście naukowym. Przykłady zaczynają się już w sformułowaniach tez rozprawy: „advanced dimensionality reduction”, „efficient identification”, „intelligent algorithm”. To, czy redukcja wymiarowości jest „advanced”, a identyfikacja „efficient”, czytelnik powinien być w stanie ocenić sam. Zgodnie zaś ze słownikiem języka polskiego, określenie algorytmu przymiotnikiem „inteligentny” wskazuje na taką cechę twórcy algorytmu.

Przechodząc do uwag merytorycznych:

1. W tekście rozprawy wielokrotnie jest mowa o poddaniu WBC działaniu promieniowania jonizującego i o zdeponowaniu w tychże komórkach dawki 1Gy. Przypuszczam (i to przypuszczenie graniczy, ale tylko graniczy, z pewnością), że zbiory oznaczone przez Doktorantkę jako A i B pochodzą z dwóch próbek krwi – tą kwestię należałoby doprecyzować, podobnie jak kwestie związane z etycznymi aspektami badania, wykorzystującego ludzki materiał tkankowy. W rozprawie nie ma w ogóle opisu układu laboratoryjnego, użytego do naświetlania komórek. Nie jest jasne, w jaki sposób ustalono, że dawka zdeponowana w WBC jest równa 1Gy. Brak standaryzacji układów laboratoryjnych używanych w eksperymentach radiobiologicznych jest ogromną przeszkodą w porównywaniu wyników uzyskanych w różnych laboratoriach stąd precyzyjny opis użytego układu jest tak ważny dla ewentualnego powtórzenia wyników, opisanych przez Doktorantkę. W rozprawie nie ma też mowy o rodzaju promieniowania jonizującego użytego w badaniach, podczas gdy na przykład 1Gy zdeponowany przez promieniowanie X daje inny efekt biologiczny (m.in. przeżywalność), niż 1Gy zdeponowany przez ciężkie jony. Ta niewiedza rzutuje z kolei na niemożność ustalenia, jakich konkretnie obszarów zastosowań promieniowania jonizującego mogą dotyczyć wyniki zaprezentowane przez Doktorantkę – czy ochrony radiologicznej personelu pracowni diagnostyki obrazowej, czy radioterapii wiązkami gamma, czy radioterapii ciężkimi jonami?

2. Na str. 15 Doktorantka, opisując algorytmy klasyfikacji, wymienia algorytmy genetyczne. Algorytmy genetyczne włącza się z reguły do klasy algorytmów optymalizacyjnych (podobnie jak np. algorytmy typu „gradient descent”, używane w treningu klasyfikatorów). W jaki sposób algorytm genetyczny może pełnić funkcję algorytmu klasyfikującego?

3. W rozdziale 4.2 (str. 25) Doktorantka opisuje algorytm zaprojektowany do selekcji cech dla klasyfikatora opartego o model regresji logistycznej. Na stronie https://scikit-learn.org/stable/modules/feature_selection.html jest opisanych szereg algorytmów selekcji cech, w tym algorytmy włączające klasyfikatory do potoku przetwarzania, algorytmy wykonujące selekcję cech w opcji „forward” i w opcji „backward” itd. Wydaje się, że algorytmy te pełnią identyczną funkcję, jak algorytm zaprojektowany przez Doktorantkę. Byłoby interesujące porównać wyniki uzyskane przez algorytm Doktorantki z algorytmami z biblioteki scikit-learn.

4. W rozprawie jest wielokrotnie mowa o „ważonych” miarach jakości klasyfikacji (np. str. 25, trzecia linia od dołu). Ważone miary jakości klasyfikacji używa się

zwykle w problemach klasyfikacji wieloklasowej. Jaka jest, na przykład, używana przez Doktorantkę definicja „ważonej” dokładności dla problemu klasyfikacji binarnej?

5. Do selekcji modeli Doktorantka używa miar w rodzaju BIC (równanie (9)) lub BF (równanie (10)). Nie jest jasne, dlaczego nie użyto po prostu walidacji krzyżowej.

6. W równaniach dostrzegam trochę nieporządku, przede wszystkim brak opisu użytych symboli (np. równanie (5) – można się tylko domyślać, czym jest $group_{cell}$, równanie (14) – indeks i pojawia się tylko po stronie prawej, w zasadzie trzeba się domyślać, że chodzi o medianę różnic itd.). Równanie (8) jest powtórzeniem równania (6).

7. Schemat treningu modelu regresji logistycznej na Rys. 3 jest nadmiarowy – to wiedza książkowa.

8. W rozdziale 5.2 przedstawione są wyniki klastrowania danych z macierzy zliczeń, po wcześniejszym zaaplikowaniu do nich analizy składowych głównych. Przekaz w tekście ostatnich akapitów staje się dla mnie trudny do uchwycenia. Z jednej strony pojawiają się oczywistości np. „It is worth emphasizing that these clusters were detected using unsupervised approach […] without providing information about cell membership”, a z drugiej niejasności np. „this analysis made it possible to conclude that the hidden data structure is related to the high heterogeneity of the data structure”. Nie rozumiem, o jaką ukrytą strukturę danych chodzi. A o heterogeniczności zbioru danych wiemy przecież od samego początku, bez spoglądania na macierze zliczeń – w końcu nasz zbiór to białe krwinki krwi, których różne typy Doktorantka wymienia w Tabeli 20. Inny przykład z tego samego rozdziału: „It can be clearly stated that the centers of clusters for control and irradiated cells, despite occupying the same separated clusters, have different locations” - dla mnie w tym zdaniu niewiele jest sformułowań „clearly stated” (jak mam na przykład rozumieć stwierdzenie, że środki klastrów zajmują te same odseparowane klastry?).

9. W rozdziale 6.1 jest napisane, że model wytrenowany na zbiorze B będzie testowany na zbiorze A. Ale z Tabeli 5 wynika, że liczba genów (cech) w zbiorze A jest o 10 mniejsza, niż liczba genów w zbiorze B. Testowanie na zbiorze A modelu wytrenowanego na zbiorze B opiera się więc na założeniu, że te 10 cech, którymi zbiory A i B się różnią, nie są istotne dla zadania klasyfikacji.

10. W równaniu (16) nie jest jasne, czym jest pozycja cechy w modelu. Można tylko przypuszczać, że jest ona związana z wartością współczynników theta, odpowiadających cesze w modelu regresji logistycznej. W opisie na str. 50 nad

i pod Rys. 50 jest więcej niejasności – np. Doktorantka używa bardzo nieprecyzyjnych sformułowań „significant difference between the following values was imperceptible” lub „without any jumps between individual values”. To czy różnice lub skoki są dostrzegalne, zależy przecież od skali, w jakiej obejrzymy Rys. 17.

11. Nie jest jasny sposób użycia równania (20) w analizach. Sumę kwadratów jakiej wielkości używa się w tym równaniu, jak wyznacza się liczbę stopni swobody?

12. Na str. 62 Doktorantka pisze: „Based on generated boxplots, individual clusters belonging to the appropriate cell subpopulation were decided”. Nie znajduję w tekście żadnego opisu sposobu przyporządkowania klastrów do typu z wykorzystaniem wykresów pudełkowych. Czy decyzja o skojarzeniu klastrów z typami WBC została podjęta w oparciu o wizualną ocenę wykresów pudełkowych?

13. Na str. 69 Doktorantka używa wyrażenia „significantly lower value of this metric”. Użycie wyrażenia „significantly lower” jest uprawnione tylko pod warunkiem wykonania odpowiedniego testu statystycznego. Testy statystyczne, służące do porównywania klasyfikatorów są opisane np. w książce N.Japkowicz i M. Shah: Evaluating Learning Algorithms. Uwaga dotyczy również innych fragmentów rozprawy np. porównania klasyfikatora opartego o regresję logistyczną z klasyfikatorem opartym o sieci neuronowe, porównań w tabeli 27, ostatni akapit na str.92 itp.

14. Na stronie 70 Doktorantka pisze, że normalizacja wartości cech jest „essential” dla klasyfikacji naświetlonych i kontrolnych komórek T. A czy był sprawdzony efekt normalizacji dla pierwotnego problemu klasyfikacji wszystkich WBC?

15. Wyniki przedstawione w Tabeli 28 uzyskano dla znormalizowanego zbioru testowego. Nie jest jasne, czym ten zbiór był normalizowany – poprawnie powinny być użyte wartości mediany i MAD dla zbioru treningowego.

16. Na str. 76 Doktorantka pisze, że celem pracy było rozpoznanie profilu genetycznego naświetlonych komórek. Nie znajduję takiego celu we wprowadzeniu. W wyniku przeprowadzonych obliczeń Doktorantka wskazała te cechy (geny), które są istotne z punktu widzenia klasyfikacji komórka naświetlona/kontrolna. Ten zbiór cech nie jest przecież tożsamy z profilem genetycznym naświetlonych komórek.

16. Rysunek 38 jest nieczytelny.

17. Metody wyjaśnialnej sztucznej inteligencji (biblioteka shap, str. 85) były użyte do oceny istotności cech w modelu sieci neuronowych. Metody te mogą być równie dobrze użyte w przypadku regresji logistycznej. Dlaczego Doktorantka nie zdecydowała się ich użyć również w tym przypadku?
18. Wykresy pokazane na Rys. 43 demonstrują dość wyraźnie, że sieci są niedotrenowane, wobec czego wyniki przedstawione w tabeli 35 mogą być zaniżone.

IV. 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.**

W swojej rozprawie Doktorantka podejmuje ważny temat dotyczący oceny wpływu promieniowania jonizującego na materiał genetyczny białych ciałek krwi. Ocena wpływu jest pośrednia – Doktorantka, w oparciu o sekwencjonowanie RNA pojedynczych komórek projektuje klasyfikatory, które umożliwiają, na podstawie wyników sekwencjonowania, z dużą dokładnością odróżnić komórki napromienione od komórek kontrolnych. Wyniki przedstawione w rozprawie przekonują mnie, że tezy rozprawy, pomijając zawarte w nich niektóre przymiotniki, zostały udowodnione tzn.:
1. „Combining feature engineering methods and dimensionality reduction techniques with unsupervised clustering algorithms allows for the identification of white blood cell subtypes in single-cell RNA sequencing data" z uwagą, że identyfikacja jest oparta również o niewymienioną w tezie wizualną ocenę pewnych wykresów, równie ważną dla ostatecznego wyniku, jak elementy

algorytmiczne, a skojarzenie klastrów z typami WBC, choć jest bardzo dobrze uargumentowane, pozostawia jednak pewna dozę niepewności, jako, że mamy do czynienia z uczeniem nienadzorowanym.

2. „The proposed algorithm of the training set construction supports the classification system, especially in the case of heterogeneous dataset” - tezę tą udowadniają wyniki uzyskane dla algorytmu selekcji cech, opartego o wielokrotne pobieranie w sposób losowy (stratified sampling) zbioru treningowego z całego zbioru danych.

Zaprojektowane narzędzia są – moim zdaniem - wymaganym przez ustawę Prawo o szkolnictwie wyższym i nauce (Art. 187 ust. 2) dla pozytywnej oceny rozprawy doktorskiej oryginalnym rozwiązaniem podjętego przez Doktorantkę problemu naukowego z obszaru dyscypliny inżynieria biomedyczna.

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 oceny wpływu promieniowania jonizującego na materiał genetyczny białych ciałek krwi.

Oceniając zatem rozprawę pozytywnie, wnioskuję o dopuszczenie Pani mgr Katarzyny Sieradzkiej 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: 2023-05-10 22:12

Zbisław Tabor