Dr hab. inż. Grzegorz Dudek, prof. PCz
Katedra Automatyki, Elektrotechniki i Optoelektroniki
Wydział Elektryczny
Politechnika Częstochowska
Al. Armii Krajowej 17
42-200 Częstochowa

Częstochowa, dn. 29 maja 2023 r.

RECENZJA

rozprawy doktorskiej mgr inż. Katarzyny Sieradzkiej
pt. *Classification of white blood cells based on single-cell sequencing data for biodosimetry purposes*

Formalną podstawą opracowania recenzji jest pismo Przewodniczącego Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej, prof. dr hab. inż. Ewy Piętki, z dnia 23.03.2023 r. Oceny rozprawy doktorskiej dokonano według kryteriów określonych w ustawie z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Promotorem rozprawy doktorskiej jest prof. dr hab. inż. Joanna Polańska, a promotorem pomocniczym jest dr Christophe Badie.

Charakterystyka rozprawy

Rozprawa napisana jest w języku angielskim, liczy 149 stron. Składa się z jedenastu rozdziałów (w tym streszczenia), bibliografii zawierającej 108 pozycji i załączników.

Tezy pracy są następujące:

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 datasets.*

W rozdziale pierwszym, *Doctoral dissertation motivation*, przedstawiono motywację, tezy i cele pracy.

Rozdział drugi, *Introduction*, omawia zagadnienia związane ściśle z tematyką rozprawy: źródła promieniowania, typy promieniowania, wpływ promieniowania na zdrowie, sekwencjonowanie RNA pojedynczych komórek, subpopulacje komórek białych krwinek, problematykę klasyfikacji komórek i przegląd literatury w zakresie selekcji cech.

W rozdziale trzecim, *Materials*, omówiono dane wykorzystywane w badaniach eksperymentalnych pozyskane na drodze sekwencjonowania pojedynczych komórek białych krwinek.

Rozdział czwarty, *Publicly available tools and developed workflows*, opisuje wykorzystywane w pracy algorytmy redukcji wymiarowości (*Uniform Manifold Approximation and Projection,* UMAP) i grupowania danych (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*; HDBSCAN) oraz przedstawia procedurę badań eksperymentalnych. W tej procedurze wykorzystuje się dwie metody uczenia maszynowego: regresję logistyczną i sieci neuronowe.

Rozdział piąty, *Preliminary data analysis*, opisuje wstępną analizę danych. Przedstawiono procedury kontroli jakości i selekcji danych (komórek i genów) oraz wyniki redukcji wymiaru i wizualizacje danych w dwuwymiarowych przestrzeniach.

W rozdziale szóstym, *Ex vivo irradiated cells' genetic profile recognition based on the logistic regression methods*, opisano procedurę badawczą mającą na celu wyznaczenie sygnatury genowej komórek napromieniowanych. Procedura ta opiera się na modelu regresji logistycznej.

W rozdziale siódmym, *Ex vivo irradiated cells' genetic profile recognition based on the neural networks*, opisano analogiczną procedurę badawczą co w rozdziale szóstym, ale opartą na sieciach neuronowych. Wyniki obu procedur porównano w rozdziale ósmym.

Badania podsumowano w rozdziale dziewiątym, a w rozdziale dziesiątym zamieszczono dyskusję. Załączniki zawierają szczegółowe wyniki badań, głównie w postaci wykresów typu boxplot.

Opinia na temat rozprawy - uwagi krytyczne i polemiczne

Sekwencjonowanie RNA pojedynczych komórek (scRNA-seq) to technika sekwencjonowania nowej generacji, która umożliwia badanie ekspresji genów na poziomie pojedynczych komórek. Zapewnia wysoką rozdzielczość badania i prowadzi do lepszego zrozumienia funkcji pojedynczej komórki w kontekście jej mikrośrodowiska. Pozwala uwzględniać inherentną heterogeniczność komórek, identyfikować ich podtypy, wyznaczać trajektorie różnicowania komórek oraz badać odpowiedzi komórek na bodźce środowiskowe lub stany chorobowe. Tematyka rozprawy dotyczy wykorzystania scRNA-seq do badania zmian ekspresji genów regulujących życie komórek pod wpływem ekspozycji na promieniowanie jonizujące. Autorka opracowała schemat przetwarzania i analizy danych genetycznych pozyskanych za pomocą scRNA-seq, obejmujący procedury selekcji, redukcji, grupowania, klasyfikacji i wizualizacji wielowymiarowych danych. Stosując ten schemat i wykorzystując nowoczesne metody analizy danych i uczenia maszynowego, zidentyfikowała geny najbardziej zaangażowane w odpowiedź na promieniowanie. Wyniki pracy są bardzo cenne. Z jednej strony pozwalają częściowo zautomatyzować pracę z wielowymiarowymi danymi genetycznymi pochodzącymi z eksperymentów sekwencjonowania pojedynczych komórek. Z drugiej strony przyczyniają się do identyfikacji pełnego profilu genetycznego komórek napromienionych. Tematykę pracy uznaję za bardzo ważną, społecznie pożyteczną i aktualną w aspekcie jej walorów poznawczych i utylitarnych.

Tytuł rozprawy jest odpowiednio zwarty i komunikatywny. W pełni oddaje najistotniejsze elementy treściowe rozprawy. Motywacja do podjęcia badań przedstawiona przez Autorkę w rozdziale pierwszym jest przekonująca. Problem badawczy jest jasno sformułowany. Cele badań, które Autorka szczegółowo uzasadnia, są czytelne: utworzenie schematu analizy bioinformatycznej w celu rozpoznania sygnatury genowej białych krwinek napromieniowanych dawką 1 Gy na podstawie danych pozyskanych za pomocą scRNA-seq oraz detekcja subpopulacji białych krwinek na podstawie tych danych. Zdefiniowano także cel pomocniczy: porównanie dwóch metod uczenia maszynowego do klasyfikacji komórek kontrolnych i napromienianych oraz identyfikacji profili genetycznych komórek napromienianych. Tezy pracy są poprawne, oryginalne i jednoznaczne, a przy tym spójne z założonymi celami.

We wstępnej części pracy (rozdział drugi) Autorka przygotowuje czytelnika do lektury kolejnych rozdziałów, definiując kluczowe zagadnienia. Charakteryzuje źródła promieniowania, typy promieniowania oraz wpływ promieniowania na zdrowie. Omawia potencjalne skutki ekspozycji

komórek na promieniowanie różnego typu i o różnej intensywności. Następnie przedstawia technikę scRNA-seq – tłumaczy jej kolejne kroki i naświetla problemy związane analizą generowanych przez nią danych genetycznych. W podrozdziale 2.5 Autorka omawia możliwości jakie daje scRNA-seq w zakresie identyfikacji podtypów komórek, w szczególności białych ciałek krwi. Charakteryzuje podtypu białych krwinek i opisuje wyniki badań dotyczących ich odpowiedzi komórkowych na promieniowanie. W kolejnym podrozdziale Autorka omawia problematykę klasyfikacji danych biologicznych, zwracając uwagę na często występujące tu problemy z wiarygodnością danych (związane z niedokładnością metod ich pozyskiwania i detekcji) oraz wielowymiarowością danych. Charakteryzuje kilka typów klasyfikatorów używanych najczęściej do danych biologicznych. Tu mam jedną uwagę – wartość przewidywaną przez k-NN otrzymuje się uśredniając (jak napisano) wartości z wybranych obserwacji w problemach regresji. W problemach klasyfikacji, a taki jest problem rozważany w pracy, stosuje się najczęściej głosowanie. W ostatnim podrozdziale przedstawiono wyniki badań literaturowych na temat selekcji cech. Podobnie jak w poprzednich podrozdziałach, ta tematyka jest zaprezentowana w bardzo czytelnej formie. Omówiono metody filtracyjne, typu *wrapper* i hybrydowe. Charakteryzując metody filtracyjne, Autorka napisała, że cechy są typowane jako istotne na podstawie ich relacji do danych wyjściowych lub korelacji z danymi wyjściowymi. Tak działają metody najprostsze. Te bardziej złożone, np. ReliefF i MRMR, uwzględniają także zależności pomiędzy cechami.

W rozdziale trzecim opisano dane wykorzystywane w eksperymentach. Wyjaśniono sposób ich wytworzenia poprzez ekspozycję białych ciałek krwi na odpowiednie promieniowanie i pozyskania danych genetycznych w procesie scRNA-seq. Skrupulatnie opisano szczegóły procedur i strukturę danych.

W rozdziale trzecim przedstawiono proponowane schematy przetwarzania danych oparte na regresji logistycznej i sieciach neuronowych. Opisano również algorytmy wykorzystane do redukcji wymiarowości (UMAP) i grupowania (HDBSCAN). Opis jest szczegółowy – podano niezbędne wzory, zdefiniowano funkcje strat, zdefiniowano metody optymalizacji i podano odpowiednie diagramy i schematy blokowe. W przypadku sieci neuronowej nie jest jednak jasne czy warstwa wyjściowa przetwarza dane liniowo, czy zawiera jakąś funkcję aktywacji, np. sigmoidę. W podsumowaniu tego rozdziału można było porównać podejścia oparte na regresji logistycznej i sieciach neuronowych, zwracając uwagę, że jedno jest liniowe, podczas gdy drugie jest nieliniowe i wyjaśniając analogie pomiędzy wzorami (3) i (11), (4) i (12) oraz (6) i (13). Schemat oparty na regresji logistycznej, który zwizualizowano na rys. 2, ma wbudowany mechanizm sekwencyjnej selekcji cech. Opis schematu opartego na sieciach neuronowych nic nie mówi o selekcji cech. Jednak z lektury rozdziału siódmego wynika, że ten schemat również wyposażony jest w mechanizm selekcji cech.

Dane pozyskane metodą scRNA-seq wymagają analizy i wstępnej obróbki. Autorka w rozdziale piątym omawia kolejne etapy obróbki danych, zaczynając od usunięcia z próbek tych komórek i genów, które nie spełniają przyjętych kryteriów jakościowych. Posługuje się przy tym swobodnie różnymi metodami selekcji danych na podstawie analizy histogramów. Jednym z takich narzędzi jest model mieszanin gaussowskich. Jednak użycie tej metody nie jest dla mnie do końca czytelne – brakuje tu definicji odpowiedniego kryterium selekcji komórek. Do redukcji danych Autorka stosuje metodę analizy głównych składowych z kryterium wyboru liczby składowych na podstawie analizy skumulowanej wariancji wyjaśnianej przez te składowe. Wyniki zwizualizowane za pomocą UMAP prowadzą Autorkę do ważnego wniosku, że separacja poszczególnych klastrów wynika nie tylko z obecności dwóch klas komórek (napromieniowanych i kontrolnych), ale przede wszystkim z dużej heterogeniczności. Ta heterogeniczność zakłóca detekcję prawidłowych profili genetycznych komórek napromieniowanych.

Rozdział szósty omawia badania dotyczące identyfikacji sygnatury genowej napromieniowanych komórek przy wykorzystaniu regresji logistycznej. Zastosowano interesujący, hybrydowy mechanizm selekcji cech, który łączy wielokrotną selekcje sekwencyjną z rankingiem cech opartym na dokładnościach klasyfikatorów zbudowanych na różnych podzbiorach cech. Autorka zauważa, że wiele genów wytypowanych tą metodą to geny odpowiedzialne za rozróżnianie poszczególnych subpopulacji komórek. Geny te „zakłócają" profil genetyczny komórek napromieniowanych. Problem rozpoznawania subpopulacji krwinek białych został podzielony na następujące etapy: selekcja cech, rozpoznawanie klastrów komórkowych za pomocą HDBSCAN oraz rozpoznawanie subpopulacji krwinek białych za pomocą genów markerowych charakterystycznych dla oczekiwanych subpopulacji. Analizując szczegółowo wyniki wizualizacji za pomocą UMAP, Autorka definiuje poprawne rozkłady subpopulacji, które zwierają nawet bardzo nieliczne subpopulacje granulocytów, komórek dendrytycznych czy bazofili/eozynofili. Wskazuje to na dużą czułość przeprowadzonej analizy. Aby uniezależnić się od heterogeniczności wynikającej z podziału na subpopulacje, Autorka w dalszej części koncentruje się na najliczniejszej subpopulacji, limfocytów T, i stosując podobny schemat badań jak wcześniej dla całej próby, identyfikuje profil genetyczny komórek napromieniowanych. Większość rozpoznanych genów odpowiedzi radiacyjnej (18 z 21) jest zgodna z doniesieniami literaturowymi. Zmniejszenie wpływu heterogeniczności poprawiło jakość klasyfikacji – uzyskano wysoką dokładność na poziomie 94%.

W rozdziale siódmym opisano badania zmierzające do identyfikacji profilu genetycznego komórek napromieniowanych oparte na alternatywnym schemacie wykorzystującym sieci neuronowe. W początkowej części tego rozdziału omówiono eksperymentalny dobór hiperparametrów. Procedura selekcji cech jest zupełnie inna niż w podejściu opartym na regresji logistycznej (ciekawy jestem dlaczego, bo hybrydowa selekcja cech tam zastosowana mogłaby się sprawdzić i w tym podejściu), wykorzystuje wartości Shapleya. W efekcie wyłoniono profil genetyczny złożony z dziesięciu genów. Większość z nich pojawia się także w profilu otrzymanym za pomocą regresji logistycznej, ale dwa to nowe geny.

Wyniki obu schematów identyfikacji profilu genetycznego komórek napromieniowanych porównano w rozdziale ósmym. Autorka z ciekawości badawczej pokusiła się jeszcze o sprawdzenie dokładności metod regresji logistycznej i sieci neuronowych na obu znalezionych profilach (21 i 10 genów). W obu przypadkach regresja logistyczna dawała większą dokładność.

W konkluzji (rozdział dziewiąty) Autorka odnosi się do tez pracy i przekonuje, odwołując się do wyników badań i zdobytej wiedzy, że zostały wykazane. Podkreśla wysoką dokładność rozpoznawania profilu genetycznego komórek napromieniowanych i podsumowuje podejścia oparte na regresji logistycznej i sieciach neuronowych. Rekomenduje model regresji logistycznej jako dokładniejszy i interpretowalny, choć w zaproponowanej wersji bardziej złożony obliczeniowo od modelu sieci neuronowych.

Rozdział dziesiąty zawiera cenną i wnikliwą dyskusję, w której Autorka omawia trudne elementy zaproponowanego schematu badań i, dostrzegając jego potencjał aplikacyjny, wskazuje sposoby jego rozszerzenia i ulepszenia.

Sekwencja treści prezentowych w kolejnych rozdziałach pracy jest właściwa: od informacji wstępnych, wyjaśniających zagadnienia poruszane w pracy, poprzez opis pozyskiwania materiału genetycznego, przygotowania i przetwarzania danych oraz charakterystykę wykorzystywanych narzędzi analizy danych i uczenia maszynowego, po omówienie proponowanych schematów badawczych, analizę

wyników i badania porównawcze. Praca napisana jest bardzo starannie i wnikliwie, poprawnym językiem naukowym z właściwym słownictwem specjalistycznym i odpowiednią ścisłością sformułowań. Układ redakcyjny rozprawy nie budzi zastrzeżeń, z jedną uwagą: zwykle dyskusję umieszcza się przed wnioskami końcowymi. Źródła literaturowe dobrane są właściwie.

Tezy rozprawy zostały udowodnione. Autorka wykazała, że stosując odpowiednie metody inżynierii cech i uczenia maszynowego można zidentyfikować podtypy białych krwinek na podstawie wyników z sekwencjonowania RNA pojedynczych komórek. Wykazała też, że opracowany schemat przetwarzania danych genetycznych pozwala uzyskać wysoką dokładność rozpoznawania komórek napromieniowanych pomimo wysokiej heterogeniczności danych. Nie mam wątpliwości, że Autorka rozwiązała postawiony problem i osiągnęła założone cele, używając właściwych metod. Oryginalność rozprawy polega na opracowaniu metodyk identyfikacji subpopulacji białych krwinek oraz rozpoznawania profilu genetycznego krwinek napromieniowanych na podstawie wyników z sekwencjonowania RNA pojedynczych komórek. Rozprawa wykorzystuje aktualne osiągnięcia w dziedzinie biologii molekularnej (metody sekwencjonowania pojedynczych komórek) i w dziedzinie analizy danych genetycznych opisane w literaturze światowej. Autorka wykazała umiejętność poprawnego i przekonywującego przedstawienia uzyskanych wyników. Na pochwałę zasługuje jej wnikliwość i drobiazgowość w opisie metod badawczych, analizie wyników badań i formułowaniu wniosków. Oceniam wysoko znaczenie uzyskanych wyników dla rozwoju dyscypliny inżynieria biomedyczna, zwłaszcza w kontekście ich potencjalnych zastosowań.

Uwagi językowe, edytorskie i redakcyjne

- Symbole zmiennych powinny być pisane konsekwentnie kursywą. W wielu miejscach pisane są czcionka prostą, np. str. 26, 27, 37, 71,
- Str. 15: opis sieci neuronowych zakończony jest cytowaniem pracy [35] z zakresu algorytmów genetycznych.
- Str. 15: zamiast „k-nn nearest neighbors" i "n-nearest neighbors" powinno być „*k*-nearest neighbors".
- Str. 15: "this algorithm predicts the membership of a given observation" – nie wiadomo o jaką przynależność chodzi.
- Str. 23: brak spacji po „[52]".
- Str. 25: zamiast „numer analyzed" powinno być „numer of analyzed".
- Str. 26: stosuje się różne symbole mnożenia – „x" i „*". Należy je ujednolicić; w wielu wzorach symbole mnożenia można pominąć.
- Str. 26: symbol logarytmu naturalnego raz pisany jest kursywą, a raz czcionka prostą (wzory (6) i (9)).
- Str. 69: brak legendy na rys. 33-35.
- Str. 71: wyjaśnienie zmiennych pod (23) jest niepotrzebne, wyjaśniono je pod (21).

Wniosek końcowy

Zakres tematyczny rozprawy doktorskiej mgr inż. Katarzyny Sieradzkiej i osiągnięte w niej oryginalne wyniki lokują tę rozprawę w obszarze inżynierii biomedycznej. Uważam, że rozprawa stanowi oryginalne rozwiązanie problemu naukowego i wskazuje na wysoki poziom wiedzy Autorki w zakresie dyscypliny inżynieria biomedyczna, a także na umiejętność samodzielnego prowadzenia przez nią badań naukowych. Pracę oceniam bardzo wysoko.

Stwierdzam, że opiniowana rozprawa doktorska spełnia wymogi ustawy z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Wnoszę o dopuszczenie mgr inż. Katarzyny Sieradzkiej do publicznej obrony pracy doktorskiej i wyróżnienie rozprawy doktorskiej.