POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań, tel. +48 61 665 2997, fax +48 61 877 1525
e-mail: office_cs@put.poznan.pl, www.put.poznan.pl

INSTYTUT
INFORMATYKI

Poznań, 2023-05-29

dr hab. inż. Agnieszka Rybarczyk
Instytut Informatyki
Politechnika Poznańska

RECENZJA ROZPRAWY DOKTORSKIEJ

mgr inż. Joanny Tobiasz

„Machine learning methods in support of multiomics signature identification for breast cancer patient subpopulations"

Promotor: prof. dr hab. inż. Joanna Polańska

Kopromotor: dr Christos Hatzis

Dyscyplina: nauki inżynieryjno-techniczne

Dyscyplina: Inżynieria biomedyczna

1. Problematyka naukowa oraz przedmiot rozprawy

Rak piersi, będący jednym z najczęściej diagnozowanych nowotworów na świecie, konsekwentnie stanowi przedmiot intensywnych badań na przestrzeni ostatnich lat. Odkrywanie mechanizmów molekularnych stojących za tym schorzeniem, identyfikowanie nowych biomarkerów oraz opracowywanie skutecznych strategii terapeutycznych to kluczowe wyzwania, które nauka stawia w tym kontekście przed bioinformatyką. Dziedzina ta, łącząca wiedzę z zakresu biologii molekularnej, genetyki, informatyki, matematyki oraz statystyki, odgrywa kluczową rolę między innymi w

przekształcaniu olbrzymich ilości danych biologicznych w użyteczną, również na gruncie medycznym, wiedzę.

Wielowymiarowe zbiory danych, generowane przez nowoczesne technologie takie jak sekwencjonowanie następnej generacji (NGS), mikromacierze, a nawet sekwencjonowanie RNA pojedynczych komórek (single-cell RNA sequencing, scRNA-seq), otwierają nowe możliwości dla naukowców. Jednakże, pozyskanie istotnych informacji z tak złożonych danych jest skomplikowane i wymaga zastosowania zaawansowanych technik ich analizy. Kluczowa rola przypada tutaj bioinformatyce, która dostarcza narzędzi i algorytmów niezbędnych do ich przetwarzania, analizowania i interpretowania.

W recenzowanej pracy podjęto z sukcesem próbę dokładniejszego scharakteryzowania podtypów nowotworu piersi przy użyciu technik uczenia maszynowego i modelowania matematycznego, jako, że choroba ta charakteryzuje się wysoce heterogenicznym obrazem molekularnym, a ustanowiona jej klasyfikacja, pozostająca niezmienna od lat, nie odzwierciedla w sposób wystarczający złożonej struktury tej choroby. Nie uwzględnia ona również profili białkowych, które jak słusznie założyła Doktorantka w swoich badaniach, mogą uzupełnić dotychczas stosowane klasyfikacje raka piersi, co ma ogromne znaczenie przy wyborze terapii. Dane proteomiczne dostarczają unikalnych informacji, które nie są dostępne na poziomie genomu czy transkryptomu. Wynika to z faktu, że białka są bezpośrednimi wykonawcami funkcji komórkowych, a ich poziomy i modyfikacje często lepiej odzwierciedlają aktualny stan komórki. Jak Doktorantka pokazała w swojej pracy, profilowanie na poziomie białek może dostarczyć bardziej kompleksowego wglądu w biologię nowotworu. Proteomiczne profile pacjentek z rakiem piersi mogą być wykorzystane do identyfikacji nowych biomarkerów prognostycznych lub predykcyjnych, co może przyczynić się do ulepszania strategii leczenia i poprawy wyników terapeutycznych.

Rozważane w pracy zagadnienia wpisują się w znaczący sposób w ostatnio intensywnie rozwijaną problematykę z zakresu integracji i analizy danych pochodzących z wysokoprzepustowych metod. Nie ulega również wątpliwości aktualność poruszanej w rozprawie tematyki.

2. Analiza treści rozprawy oraz uzyskanych wyników

2.1. Treść rozprawy

Praca została napisana w języku angielskim, ma 160 stron i składa się z 8 rozdziałów, spisu tabel, rysunków, skrótów, źródeł finansowania oraz obszernego wykazu literatury. Praca posiada zwartą konstrukcję i dobrze oddaje kolejne etapy realizacji przyjętego w pracy celu.

Rozdział 1 poprzedzony jest streszczeniem w języku polskim („Streszczenie") oraz streszczeniem w języku angielskim („Abstract"). W rozdziale 1 („Introduction") Autorka określa kontekst tematyki rozprawy, cel i zakres pracy oraz formułuje tezy, które zostały przebadane w pracy. Doktorantka postawiła sobie za cel zastosowanie metod uczenia maszynowego do identyfikacji i klinicznego oraz molekularnego scharakteryzowania podpopulacji pacjentek z nowotworem piersi. Jednocześnie poprzez odpowiednio dobrane kompleksowe metody analizy statystycznej wsparte przez miarę wielkości efektu zmierzyła się z problemem niezbalansowanych, zróżnicowanych pod względem wielkości grup.

W kolejnym, 2 rozdziale Autorka nakreśla tło biologiczne rozprawy, szczegółowo omawia kliniczne i molekularne klasyfikacje raka piersi wraz z ich ograniczeniami, rzeczowo i przekonująco uzasadniając potrzebę dalszych badań w kierunku bardziej zgodnej i dokładniejszej ich charakterystyki. Przedstawia również wybrane, najistotniejsze podejścia stosowane do wyodrębniania podtypów nowotworu piersi.

W rozdziale 3 Autorka zawarła opis danych rzeczywistych pochodzących z projektu TCGA-BRCA (ang. *The Cancer Genome Atlas Program - Breast Invasive Carcinoma*) oraz bardzo szczegółowo przedstawiła metody użyte do analizy danych proteomicznych, takie jak: algorytmy klastrowania, metody inżynierii cech, miary i testy statystyczne, przyjmując porządek opisu zgodny z ich wykorzystaniem na poszczególnych etapach przeprowadzonych w rozprawie analiz.

Rozdziały 4, 5 i 6 zawierają główne, oryginalne osiągnięcia Autorki rozprawy.

I tak: rozdział 4 przedstawia zaprojektowane przez Doktorantkę podejście dedykowane do wyodrębnienia jednoznacznych podpopulacji pacjentek w oparciu o profil białkowy. Autorka przeprowadziła analizę i ocenę różnych kombinacji metod inżynierii cech i klastrowania, wykazując,

że algorytm DiviK (rozwijany w zespole prof. Polańskiej) osiągnął lepsze rezultaty od pozostałych metod. Na podstawie jego wyników zidentyfikowano sześć podtypów raka piersi: podstawny, HER2-wzbogacony, luminalny B oraz trzy podgrupy luminalne A: A1, A2 i A3. Sprawdzono również dane po kątem wystąpienia tzw. efektu paczki (ang. *batch effect*, systematycznego efektu specyficznego dla zestawu danych, niewynikającego ze zmienności o podłożu biologicznym) oraz przeprowadzono jego korektę.

W rozdziale 5 Autorka dokonała oceny sześciu, wyodrębnionych na wcześniejszym etapie analizy, podpopulacji pacjentek, pod kątem demograficznym i klinicznym. Potwierdziła różnice w przeżyciu między zdefiniowanymi podpopulacjami, także w ramach nowo odkrytych podgrup luminalnych. Autorka stwierdziła także niewielkie powiązanie między badanymi podtypami raka a czynnikami demograficznymi lub klinicznymi, podobnie jak ma to miejsce w przypadku podtypów opartych na teście PAM50. Wykryła również, że zidentyfikowane podpopulacje wykazują zróżnicowanie we frakcjach komórek immunologicznych.

Rozdział 6 poświęcony jest wykrywaniu markerów specyficznych dla wyodrębnionych podpopulacji oraz sygnatur pozwalających na ich rozróżnienie. Dzięki zastosowaniu technik uczenia maszynowego, udało się odkryć specyficzną sygnaturę białkową, która pozwala na rozróżnienie wszystkich sześciu wykrytych podtypów. Wykazano również, że nowe podtypy luminalne wykazują mniejsze zróżnicowanie na poziomie transkryptomu w porównaniu do zróżnicowania na poziomie proteomu.

Rozdział 7 stanowi udane podsumowanie rozprawy.

Rozdział 8, pełniący rolę dodatku do pracy, zawiera kompleksowy zbiór dodatkowych tabel i ilustracji wynikających z przeprowadzonych analiz.

2.2. Najważniejsze wyniki przedstawione w rozprawie

Za najważniejsze wyniki przedstawione w rozprawie uznać można:

1. Opracowanie zaawansowanego procesu analizy wielowymiarowych, wysokoprzepustowych danych z wykorzystaniem metod uczenia maszynowego w celu identyfikacji podpopulacji raka piersi, a także ich klinicznych i molekularnych charakterystyk.
2. Zaproponowanie i uzasadnienie (na podstawie przeprowadzonych analiz opartych na rzeczywistych danych) następujących hipotez, o kluczowym znaczeniu dla naukowców realizujących badania w oparciu o dane wysokoprzepustowe w analizie raka piersi:
    - Zastosowanie metod uczenia maszynowego oraz modelowania matematycznego pozwoliło na zidentyfikowanie nowych, molekularnie różniących się podpopulacji pacjentek z nowotworem piersi.
    - W sytuacji, gdy mamy do czynienia z grupami o zróżnicowanym rozmiarze, odpowiednio dobrane testy statystyczne, wsparte miarą wielkości efektu, pozwalają na precyzyjne określenie profili podtypów na poziomie molekularnym oraz klinicznym.
3. Uzyskanie znaczących wyników biologicznych oraz:
    - Pokazanie, że dane proteomiczne mają ogromny potencjał w identyfikacji podpopulacji pacjentek z rakiem piersi.
    - Zidentyfikowanie 6 podtypów raka piersi, w tym dodatkowego podziału w obrębie grupy luminalnej A.
    - Wykazanie, że zidentyfikowane podpopulacje różnią się pod względem wyników klinicznych i molekularnych, co sugeruje możliwość indywidualizacji terapii.
    - Uzyskanie zestawu markerów specyficznych dla wyodrębnionych subpopulacji oraz sygnatury białkowej pozwalającej na rozróżnienie wszystkich podtypów.

2.3. Uwagi dyskusyjne

Praca jest poprawnie napisana i zasługuje na wysoką ocenę merytoryczną. Poniższe uwagi mają głównie charakter dyskusyjny. Oto niektóre z nich:

- Niektóre rysunki są trudne do zinterpretowania, ze względu na wykorzystanie przez Doktorantkę kolorów o zbyt podobnej tonacji (np. Rysunek 4.2, 4.3 (dla podtypu luminalnego), 5.1, 6.1).
- Rozprawa zawiera wyczerpujące wprowadzenie z zakresu zagadnień dotyczących algorytmów klastrowania, metod inżynierii cech, miar i testów statystycznych. Jednakże wstęp dotyczący pojęć z zakresu biologii molekularnej oraz metod pozyskiwania danych wysokoprzepustowych jest dość pobieżny.
- W kontekście badań w zakresie współzależności pomiędzy zmiennymi (Strona 45), czy rozważała Pani zastosowanie zamiast testu niezależności chi-kwadrat Pearsona z poprawką Yatesa, innego testu np. dokładnego testu Fishera?
- Sekwencjonowanie RNA pojedynczych komórek (scRNA-seq) jest nowoczesną metodą, która jest szczególnie przydatna do wykrywania heterogeniczności na poziomie komórkowym w obrębie próbki, co jest często spotykane w przypadku nowotworów, np. takich jak rak piersi. Pozwala to na identyfikację subpopulacji komórek, które mogą różnić się pod względem swojego potencjału patogennego, zdolności do tworzenia przerzutów czy odporności na terapie. Jestem ciekawa zdania Pani na temat wykorzystania tego typu danych do identyfikowania i molekularnej charakterystyki subpopulacji pacjentek z rakiem piersi.
- Projekty badawcze, które szeroko wykorzystują sekwencjonowanie i analizę ekspresji genów w kontekście onkologii, często komplementują te metody o zbieranie danych z ankiet, nierzadko obejmujących informacje o historycznej zachorowalności na danego raka w rodzinie, z wyszczególnieniem wieku, relacji pokrewieństwa czy podtypu badanego nowotworu. Takie dane mogą posłużyć do stworzenia np. drzew rodowodowych. Ciekawa jestem jak Pani widzi potencjał takich danych w kontekście pogłębienia naszego rozumienia przeżywalności w odniesieniu do różnych podtypów raka piersi. Jak proponowałaby Pani uwzględnienie takich danych w analizie i jakimi metodami?

2.4. Uwagi redakcyjne

Nie będąc native speakerem recenzent nie ocenia stylu przedłożonej rozprawy. Zasługuje na uwagę jednak przejrzystość opisu oraz konsekwencja w użyciu dobrze zdefiniowanych pojęć. W zrozumiały sposób przedstawione są zastosowane wskaźniki, formuły obliczeniowe oraz cała procedura badawcza. Praca napisana jest precyzyjnym językiem i starannie zredagowana, mimo że Autorka nie ustrzegła się drobnych błędów. Wymienię niektóre z nich:

- str. 8 (...) at the end of 2020, 7.8 females alive were diagnosed with breast cancer [powinno być: at the end of 2020, 7.8 million females who are currently alive had been diagnosed with breast cancer]
- str. 11 (...) Initially, the within-subtype diversity in clinical outcomes was assumed toe reflected [powinno być: Initially, the diversity in clinical outcomes within each subtype was assumed to be reflected]
- str. 13 (...) Chemotherapy is a man treatment option [powinno być: Chemotherapy is a main treatment option]
- str. 36 (...) where *k*was [powinno być: where *k* was]
- str. 42 (...) the same importance in put on differences [powinno być: the same importance is put on differences]

3. Podsumowanie i konkluzja oceny

Praca pokazuje szerokie spektrum zainteresowań autorki oraz szeroką wiedzę przedmiotu. Autorka pokazała imponującą zdolność do poruszania się w skomplikowanych dziedzinach, takich jak analiza proteomiczna, uczenie maszynowe i statystyka. Zastosowanie nowatorskich metod obejmujących kombinacje różnych algorytmów w zakresie inżynierii cech i klastrowania dowodzi zaawansowanego zrozumienia dla tych technik i ich potencjalnej mocy w kontekście badania raka piersi. Przedstawiona do oceny rozprawa zawiera oryginalne i wartościowe wyniki, które stanowią istotny wkład w zrozumienie heterogeniczności raka piersi na poziomie proteomicznym. Rezultaty te, uzyskane dzięki zastosowaniu technik uczenia maszynowego i matematycznego modelowania, otwierają nowe perspektywy dla indywidualizacji terapii i potencjalnej optymalizacji planowania leczenia. Ponadto, identyfikacja nowych

markerów i podpopulacji pacjentów może przyczynić się do dalszych badań nad nowymi opcjami terapeutycznymi.

Wymienione wcześniej uwagi dyskusyjne i drobne zastrzeżenia nie umniejszają osiągnięć Doktorantki i nie mają istotnego wpływu na wagę oraz jakość przedstawionych w pracy wyników. Reasumując można stwierdzić, że główne wyniki potwierdzają osiągnięcie z powodzeniem założonego w pracy celu. Zastosowane przez autorkę metody badawcze są właściwe dla podjętej przez nią problematyki, a dobór cytowanej literatury nie budzi zastrzeżeń. Podsumowując, należy także stwierdzić, że przedstawiona do oceny rozprawa zawiera oryginalne i wartościowe wyniki naukowe, które stanowią istotny wkład w dziedzinę nauki jaką jest bioinformatyka.

W związku z powyższym, stwierdzam, że praca pt. „Machine learning methods in suport of multiomics signature identification for breast cancer patient subpopulations" spełnia wymagania stawiane rozprawom doktorskim przez Ustawę o stopniach naukowych i tytule naukowym, i w konsekwencji może stać się przedmiotem publicznej obrony. Wnoszę zatem o dopuszczenie mgr inż. Joanny Tobiasz do dalszych etapów postępowania o nadanie stopnia doktora.

Podpisała Agnieszka Rybarczyk (podpis odręczny)