Wydział Automatyki, Elektroniki i Informatyki
Katedra Inżynierii i Analizy Eksploracyjnej Danych

Prof. dr hab. inż.
Joanna Polańska
Kierownik Katedry

Recenzja rozprawy doktorskiej

Autor: Mgr inż. Alicja Płuciennik

Tytuł: Algorytmy integracji danych i uczenia maszynowego w diagnostyce nowotworów

Promotor: prof. dr hab. inż. Krzysztof Fujarewicz

Promotor pomocniczy: dr hab. Małgorzata Oczko-Wojciechowska

Opiekun przemysłowy: dr inż. Michał Bachorz

Dziedzina: Nauki inżynieryjno-techniczne

Dyscyplina: Inżynieria biomedyczna

Niniejsza recenzja została przygotowana na zlecenie Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej (pismo RDIB.002.49.2022) z dnia 23.09.2022 roku.

Ogólna charakterystyka rozprawy i ocena wyboru tematyki rozprawy

Przedłożona do recenzji rozprawa doktorska liczy 104 strony i zawiera 13 części: spis treści, 6 rozdziałów, bibliografię, streszczenie i jego tłumaczenie na język angielski, spis rysunków, spis tabel oraz wykaz skrótów. Bibliografia to 80 pozycji, głównie artykuły z ostatnich lat. W dysertacji nie umieszczono oddzielnego wykazu osiągnięć Doktorantki, co trochę utrudnia ocenę dorobku naukowego. Kolejność rozdziałów: (1) motywacja, (2) cel pracy, (3) materiał i metody, (4) wyniki, (5) dyskusja oraz (6) wdrożenie, jest logiczna i w pełni oddająca istotę przeprowadzonych badań.

Tematyka projektu doktorskiego jest związana z obszarem badawczym grupy naukowej, której członkinią jest Doktorantka i skupia się na poszukiwaniu markerów molekularnych choroby

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 2144 / +48 601 891 245
joanna.polanska@polsl.pl

nowotworowej, a w szczególności raków tarczycy oraz piersi. Wsparcie klasycznych metod diagnostycznych testami molekularnymi stanowi istotną zmianę jakościową w praktyce klinicznej i pozwala na skuteczne wdrożenie personalizacji terapii, której celem jest maksymalizacja skuteczności przy minimalizacji toksyczności leczenia, co w konsekwencji prowadzi do zmniejszenia kosztów społecznych choroby. Przykładem jest tutaj rak piersi, w którym już jakiś czas temu, na podstawie profili transkryptomicznych, zidentyfikowano markery molekularne rozróżniające jego podstawowe podtypy. Dla każdego z nich opracowano odmienne, dopasowane do specyfiki choroby, protokoły leczenia zmniejszając istotnie śmiertelność oraz częstość występowania efektów ubocznych wśród pacjentek. Jednak zadanie poszukiwania markerów molekularnych i tym samym przeprowadzania jak najmniej inwazyjnej, ale będącej równocześnie precyzyjną, diagnostyki nie jest jeszcze ukończone. Istniejące rozwiązania nadal nie wykrywają szeregu podtypów raka, a dla niektórych nie udało się jeszcze znaleźć skutecznego markera przesiewowego lub diagnostycznego. Za jedną z przyczyn uznaje się słabą generalizację bazujących na biomarkerach opracowanych modeli predykcyjnych, co spowodowane być może, między innymi, zawężeniem zbioru uczącego do jednej grupy pacjentek/pacjentów. Duże nadzieje pokłada się też w badaniach multiomicznych, gdzie dzięki kombinacji pomiarów klinicznych, transkryptomicznych, protomicznych czy obrazowych, lepiej reprezentowane są procesy komórkowe, zaburzone w trakcie nowotworzenia. Niezależnie od tego, na którym aspekcie z tych dwóch wymienionych powyżej skupiać się będą w najbliższej przyszłości badania naukowców, integracja danych jest krokiem niezbędnym i nieuniknionym. Dlatego tematykę badań podjętą w przedstawionej rozprawie doktorskiej uważam za niezwykle istotną i bez wątpienia wpisującą się w obszar inżynierii biomedycznej. Potencjał wdrożeniowy wyników jest wysoki a same badania uważam za kluczowe dla onkologii oraz biologii molekularnej.

W pracy postawiono dwie tezy badawcze:

1. *Fuzja markerów molekularnych oraz danych cytologicznych obecnie wykorzystywanych w trakcie diagnostyki pacjenta poprawia dokładność klasyfikacji guzów o charakterze złośliwym lub*

łagodnym w przypadku dobrze zróżnicowanych raków tarczycy lub zmniejsza liczbę potrzebnych cech molekularnych koniecznych do osiągnięcia równie dobrej jakości.

2. *Fuzja markerów molekularnych oraz markerów immunohistochemicznych wykorzystywanych w trakcie diagnostyki pacjentek z nowotworami piersi poprawia dokładność klasyfikacji podtypu choroby lub zmniejsza liczbę potrzebnych cech molekularnych koniecznych do osiągnięcia równie dobrej jakości.*

oraz zdefiniowano jeden cel wdrożeniowy:

1. *Celem komercyjnym projektu doktoratu wdrożeniowego było wybranie algorytmów wykorzystujących integrację danych molekularnych i klinicznych w celu podniesienia jakości klasyfikacji pacjentów, przy jednoczesnym obniżeniu kosztów badań molekularnych.*

Omówienie rozprawy

Poniżej pokrótce przedstawię podsumowanie poszczególnych części rozprawy doktorskiej. W rozdziale 1, zatytułowanym „Motywacja", Doktorantka przedstawiła podstawowe informacje dotyczące zapadalności i umieralności na choroby nowotworowe, skupiając się na nowotworach piersi i tarczycy. Omówiła również aktualnie obowiązujące protokoły diagnostyczne tych chorób, wskazując na ich silne i słabe strony. Ostatni z podrozdziałów poświęcono przedstawieniu możliwych do zastosowania procedur integracji i fuzji danych. W swojej pracy Doktorantka słusznie rozróżnia pojęcia integracji i fuzji danych, definiując tę drugą jako proces łączenia danych z różnych sensorów. Przedstawia kilka możliwych podejść do procesu integracji oraz fuzji danych podając przykłady z obszaru badań na nowotworami piersi i tarczycy. Rozdział 2 pracy to przede wszystkim omówienie celu i przedstawienie hipotez badawczych będących podstawą dysertacji. W całej rozprawie, w tym również w tym rozdziale, pojawia się wiele odniesień do wyników prac realizowanych w ramach projektu MILESTONE – nie zawsze jest jasne co jest wynikiem badań Doktorantki a co reszty osób zaangażowanych w te badania. Rozdział 3, nazwany „Materiał i metody", zaczyna się od przedstawienia zbiorów danych wykorzystywanych na etapie weryfikacji postawionych hipotez. W przypadku raka

tarczycy materiał bazowy to profile transkryptomiczne 200 dawców (77 osób z rozpoznaną chorobą nowotworową oraz 123 osoby ze zmianami łagodnymi). Nie posługiwano się tutaj informacją o pełnym transkryptomie, zbiór danych zawężono do dwóch podzbiorów cech: 40 oraz 163 genów zidentyfikowanych we wcześniejszych badaniach jako skuteczne sygnatury raka tarczycy i jego podtypów (wniosek patentowy zespołu, którego członkinią jest Doktorantka). Zbiór dodatkowy to tzw. dane kliniczne, w tym wiek, płeć, wielkość guza oraz wynik oceny cytologicznej. Drugi zestaw danych dotyczył chorych na raka piersi typu luminalnego wg klasyfikatora PAM50. Dostępna była dla każdej z pacjentek informacja o ich profilu proteomicznym, mierzonym ilościowymi technikami spektrometrii mas (dla 172 białek), oraz o poziomie aktywności 50 markerów transktyptomicznych stanowiących przestrzeń cech dla klasyfikatora PAM50. Do analizy włączono też informacje o wieku pacjentek w momencie rozpoznania choroby, wynikach badań immunohistochemicznych (dla receptorów estrogenowego ER, progesteronowego PR oraz HER2) oraz wartości TNM. Dla uproszczenia skupiono się na zadaniu identyfikacji dwóch podtypów nowotworów: luminalnego A (166 próbek) i B (98 próbek).

W obu przypadkach, w dalszych badaniach nie posługiwano się w zasadzie podstawowymi danymi klinicznymi a ich agregatem, określanym mianem ryzyka klinicznego. Ryzyko kliniczne jest nową cechą wyznaczaną z wykorzystaniem sieci bayesowskich na podstawie wymienionych powyżej cech klinicznych i została zaproponowana jako efekt pracy większego zespołu badawczego z projektu MILESTONE.

Rozdział 3 zawiera również omówienie metod wykorzystywanych w dalszych analizach, w tym analizy zależności pomiędzy zmiennymi oraz ogólnego schematu procesu uczenia i walidacji klasyfikatorów. Przedstawiono tutaj także szczegółowe scenariusze dwóch testowanych schematów fuzji danych: *wczesnej* (na etapie selekcji cech) oraz *późnej* (po etapie selekcji cech ze zbioru podstawowego) oraz wprowadzono pojęcia dotyczące rankingu cech, stabilności sygnatury czy metod oceny modeli.

Rozdział 4 „Wyniki" to główny element rozprawy doktorskiej, w którym przedstawiono wyniki poszczególnych analiz. Rozpoczyna się od wizualizacji zależności pomiędzy cechami, wskazując na ich wysokie powiązanie ze stanem klinicznym i możliwą redundancję. Przeprowadzone testy

statystyczne potwierdziły przytoczone w rozdziale 3 wnioski z wcześniejszych badań o relatywnie niewielkie mocy dyskryminacyjnej pojedynczych cech z profilu klinicznego pacjentów. W dalszej części rozdziału umieszczono wyniki klasyfikacji dla oryginalnych zbiorów danych oraz połączonych z danymi klinicznymi według dwóch wybranych scenariuszy fuzji. Badania te przeprowadzono dla obu zbiorów danych, przy czym w przypadku raka piersi analizę przeprowadzono niezależnie dla połączenia proteomiki oraz transkryptomiki. Uzyskane sygnatury poddano następnie szczegółowej analizie ich stabilności a wyniki jakości predykcji - statystycznej analizie porównawczej.

W rozdziale 5 przedstawiono podsumowanie całości badań i odniesiono je do postawionych na początku rozprawy hipotez badawczych i szczegółowych celów. Z kolei rozdział 6, niezwykle interesujący, pokazuje potencjał zastosowań omawianych technik w praktyce, przy czym zauważyć tutaj można wzbogacenie opracowanego potoku konstrukcji systemów o etap optymalizacji struktury klasyfikatora (rozumiany tutaj jako wprowadzenie kryterium stopu do procesu krokowej rozbudowy modelu).

W trakcie lektury dysertacji nasunęło mi się kilka pytań, które przedstawiam poniżej:

1) Na stronie 37, w rozdziale dotyczącym stabilności selekcji pojawia się zdanie: *„Wysoka stabilność [selekcji] świadczy o odporności algorytmu na zmiany w składzie próbek w zbiorze uczącym, co świadczy o lepszej generalizacji modelu, niż w przypadku zastosowania metody selekcji o niskiej stabilności."* Czy obserwowana stabilność selekcji cech (lub jej brak) jest jedynie własnością algorytmu selekcji cech? Czy często występująca w rzeczywistych danych heterogeniczność wewnątrzklasowa (rozumiana jako występowanie podtypów niezdefiniowanych w procesie formułowania zadania klasyfikacji) może mieć swoje odzwierciedlenie w stabilności selekcji? Chętnie poznałabym opinię Doktorantki w tym zakresie.
2) Na wykresie 4.1, panel *„Raki luminalne, zestaw cech PAM50"* obserwujemy wyraźną bimodalność rozkładu współczynników korelacji. Czy mogę prosić o interpretację tego zjawiska?
3) Rysunek 4.3 przedstawia dystrybucję wyników testów Kruskala-Wallisa. W opisie metod podano, że wyniki testów były poddane korekcie Benjaminiego-Hochberga z powodu wielokrotnego

testowania. Czy rysunek przedstawia surowe wartości p z testów czy też wartości FDR z korekty Benjaminiego-Hochberga?

4) Na str. 59 pojawia się stwierdzenie: *„W przypadku modeli klasyfikujących guzy tarczycy z ryzykiem złośliwości obserwujemy poprawę dokładności klasyfikacji już dla niewielkiej liczby cech".* Czy mogę prosić o przybliżenie podstawy do takiego wniosku? Z załączonych wykresów wynika, że z drobnymi wyjątkami przedziały ufności dla bardzo szerokiego zakresu liczby cech się pokrywają. Mam również drugie pytanie/uwagę. Czy wpływu fuzji danych (szczególnie wg scenariusza tzw. *późnej fuzji)* nie powinno się na etapie porównań przeprowadzać jako wyniki dla pierwotnej liczby cech versus wyniki dla pierwotnej + cechy z fuzji? Oznaczałoby to, jeśli mamy 4 cechy kliniczne, porównanie przykładowo oryginalnego modelu 6 cech z modelem 6+4=10 cech – widać to bardzo np. na rys. 4.16 panel górny. Czym kierowała się Doktorantka wybierając przedstawiony w pracy schemat porównań.
5) Nie jest jasne dlaczego nie przeprowadzono drugiego poziomu optymalizacji modelu (tzn. dla każdego zadania nie wybierano modelu o optymalnej liczbie cech według jakiegoś wskaźnika jakości) a jedynie analizowano modele o zadanej a priori liczbie cech (wyniki na rysunkach 4.12-4.15 oraz 4.16-4.19). Taki proces wykonano już w badaniach przedstawianych w rozdziale 6.
6) Dlaczego nie zdecydowano się w przypadku analizy danych dotyczących raka piersi na połączenie danych proteomicznych, transkrytomicznych oraz klinicznych?
7) Nie do końca rozumiem hipotezę postawioną w rozdziale 4.4. Czy mogę prosić o więcej szczegółów?

Ocena rozprawy

Moja ogólna ocena rozprawy jest **pozytywna**. Poniżej przedstawione są jej poszczególne elementy.

Pierwszym elementem oceny jest rozważenie indywidualnego wkładu Doktorantki w badania naukowe przedstawione w rozprawie, w aspekcie faktu, że pracuje ona w wieloosobowym zespole

badawczym, była jedną z wykonawczyń projektu MILESTONE oraz, że wszystkie te publikacje są wieloautorskie. Nie załączono do rozprawy wykazu publikacji Doktorantki, posłużę się więc informacjami jakie znalazłam w bibliografii dysertacji oraz na stronach odpowiednich baz danych:

Publikacje w recenzowanych czasopismach naukowych o niezerowym współczynniku wpływu:

1. **Płuciennik A.**, Płaczek A., Wilk A., Student S., Oczko-Wojciechowska M., Fujarewicz K.: *Data integration - possibilities of molecular and clinical data fusion on the example of thyroid cancer diagnostics,* International Journal of Molecular Sciences, MDPI, vol. 23, nr 19, 2022, Numer artykułu: 11880, s. 1-16, DOI:10.3390/ijms231911880, IF=6,208
2. Płaczek A., **Płuciennik A.**, Kotecka-Blicharz A., Jarząb M., Mrozek D.: *Bayesian assessment of diagnostic strategy for a thyroid nodule involving a combination of clinical synthetic features and molecular data,* IEEE Access, vol. 8, 2020, s. 175125-175139, DOI:10.1109/ACCESS.2020.3026315, IF=3,367
3. **Płuciennik A.**, Stolarczyk M., Bzówka M., Raczyńska A., Magdziarz T., Góra A.: *BALCONY: an R package for MSA and functional compartments of protein variability analysis,* BMC Bioinformatics, vol. 19, nr 1, 2018, Numer artykułu: 300, s. 1-8, DOI:10.1186/s12859-018-2294-z, IF=2,511
4. Magdziarz T., Mitusińska K., Gołdowska S., **Płuciennik A.**, Stolarczyk M., Ługowska M., Góra A.: *AQUA-DUCT: a ligands tracking tool,* Bioinformatics, vol. 33, nr 13, 2017, s. 2045-2046, DOI:10.1093/bioinformatics/btx125, IF=5,481

Publikacje w recenzowanych materiałach konferencji naukowych:

5. Student S., **Płuciennik A.**, Łakomiec K., Wilk A., Bensz W., Fujarewicz K.: *Integration strategies of cross-platform microarray data sets in multiclass classification problem*, W: Computational science and its applications: **ICCSA 2019. 19th International conference, Saint Petersburg, Russia**, July 1-4, 2019. Proceedings / Misra S.[i in.](red.), Lecture Notes In Computer Science, vol. 11623, 2019, Springer, ISBN 978-3-030-24307-4, s. 602-612, DOI:10.1007/978-3-030-24308-1_48
6. Student S., Łakomiec K., **Płuciennik A.**, Bensz W., Fujarewicz K.: *Classification system for multi-class biomedical data that allows different data fusion strategies*, W: Information technology in

biomedicine: **International Conference, ITIB 2019, Kamień Śląski, Poland**, June 18-20, 2019 / Piętka Ewa[i in.](red.), Advances in Intelligent Systems and Computing, vol. 1011, 2019, Springer, ISBN 978-3-030-23761-5, s. 593-602, DOI:10.1007/978-3-030-23762-2_52

7. Płaczek A., **Płuciennik A.**, Pach M., Jarząb M., Mrozek D.: *The role of feature selection in text mining in the process of discovering missing clinical annotations - case study*, W: Beyond databases, architectures and structures: Paving the road to smart data processing and analysis. **15th International conference, BDAS 2019, Ustroń, Poland**, May 28-31, 2019. Proceedings / Kozielski Stanisław[i in.](red.), Communications in Computer and Information Science, vol. 1018, 2019, Springer, ISBN 978-3-030-19092-7, s. 248-262, DOI:10.1007/978-3-030-19093-4_19
8. Student S., **Płuciennik A.**, Jakubczak M., Fujarewicz K.: *Feature selection based on logistic regression for 2-class classification of multidimensional molecular data*, W: Artificial intelligence: methodology, systems, and applications: **18th International Conference, AIMSA 2018, Varna**, Bulgaria, September 12-14, 2018. Proceedings / Agre G., van Genabith J., Declerck T. (red.), Lecture Notes In Computer Science, vol. 11089, 2018, Springer, ISBN 978-3-319-99343-0, s. 286-290, DOI:10.1007/978-3-319-99344-7_29

Streszczenia wystąpień konferencyjnych:

9. **Płuciennik A.**, Płaczek A., Łakomiec K., Fujarewicz K.: *Wpływ niezbalansowania danych na problem klasyfikacji w prospektywnym badaniu raka tarczycy*, W: **V Śląskie Spotkania Naukowe, Bobolice, 25-26 Maja 2018 r**, 2018, [b.w.], s. 13-14
10. **Płuciennik A.**, Pacholczyk M: *Induced fit docking with AutoDock and Modeller*, W: **XI Symposium of Polish Bioinformatics Society, September 5-7, 2018, Wrocław,** Poland: Book of abstracts, 2018, [b.w.], s. 66
11. Gołdowska S., Markowska K., **Płuciennik A.**, Góra A.: *Modification of the human soluble epoxide hydrolase active site accessibility by engineering of its entrance tunnel*, W: Chemistry towards biology : **8th Central Europe Conference, Brno, Czech Republic, 28th August - 1st September**

2016. Book of abstracts, 2016, University of Veterinary and Pharmaceutical Sciences, ISBN 978-80-7305-777-0, s. 97

12. **Płuciennik A.**, Stolarczyk M., Magdziarz T., Góra A.: *Balcony - better alignment consensus analysis*, W: **XIXth Gliwice Scientific Meetings 2015, Gliwice, November 20-21, 2015** [online], 2015, [b.w.], (plik pdf) 130

Za główną publikację związaną z ocenianą rozprawą doktorską należy uznać pracę oznaczoną numerem [1], w której Doktorantka jest pierwszym autorem. Związek z rozprawą mają również, w mojej opinii, prace [2], [5], [6] oraz [8], przy czym praca [2] poświęcona jest budowie modelu oszacowania ryzyka klinicznego a ten temat wydaje się być przedmiotem rozprawy doktorskiej innego członka zespołu MILESTONE. Z lektury artykułów będących dorobkiem naukowym Doktorantki oraz dysertacji wyłania się solidny i różnorodny warsztat naukowy oraz duży udział w zrealizowaniu wartościowych i oryginalnych projektów naukowych. Widać także kompetencje Doktorantki oraz dobre panowanie nad całokształtem złożonych aspektów naukowych związanych z omawianymi publikacjami.

Drugim aspektem oceny jest ranga, jakość i oryginalność wyników naukowych zawartych w publikacjach wchodzących w skład rozprawy. Rangę, jakość i oryginalność wyników należy także ocenić pozytywnie. Dwie prace zostały opublikowane w bardzo renomowanych czasopismach naukowych. Ponadto należy podkreślić spójność tematyczną wspomnianych powyżej pięciu publikacji. Wszystkie prace bazują na oryginalnych zbiorach danych eksperymentalnych (obserwacyjnych) oraz klinicznych. Wyniki w nich umieszczone uzupełniają się nawzajem i jako całość tworzą logiczny i uporządkowany zbiór tez naukowych, które wzbogacają teorię konstrukcji modeli w zadaniach uczenia maszynowego.

Wreszcie jako trzeci aspekt oceny warto uwzględnić dorobek naukowy Doktorantki, który nie mieści się w pięciu wspomnianych powyżej publikacjach – chodzi o prace [3], [4] oraz [7]. Należy stwierdzić, że jest to dorobek doktorski, który ma bardzo znaczną wielkość i jakość naukową (publikacje w czasopismach *Bioinformatics* oraz *BMC Bioinformatics* – oba czasopisma zaliczane do grupy co najmniej top10 wg bazy SCOPUS). Tematyka tych prac jest bardzo różnorodna, zogniskowana

jednak na zagadnieniach współczesnej bioinformatyki oraz onkologii eksperymentalnej, klinicznej, a także częściowo obliczeniowej.

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

1) Określenie zintegrowanego, multiomicznego profilu raka tarczycy i oszacowanie jego mocy dyskryminacyjnej pomiędzy podtypami.
2) Wykazanie skuteczności fuzji informacji klinicznej i molekularnej na przykładzie predykcji podtypów raka piersi.
3) Opracowanie potoku przetwarzania informacji i optymalizacji struktury klasyfikatora.

Poprawność treści rozprawy nie wzbudza zastrzeżeń, a stwierdzenia w niej zawarte wydają się być w pełni godne zaufania, co wynika w szczególności z przedstawionych podstaw teoretycznych popartych wynikami przeprowadzonych badań eksperymentalnych.

Syntetyczna o cena rozprawy:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? Zdecydowanie TAK

B. Czy kandydatka posiada ogólna wiedzę teoretyczną w dyscyplinie? Zdecydowanie TAK

C. Czy posiada umiejętność samodzielnego prowadzenia pracy naukowej? Zdecydowanie TAK

Konkluzja

Osiągnięcia i oryginalne elementy publikacji wchodzących w skład rozprawy, a także jakość całego tekstu rozprawy, są na pewno wystarczające do jej ogólnej pozytywnej oceny oraz spełniają zwyczajowe wymagania stawiane rozprawom doktorskim.

Stwierdzam zatem z pełnym przekonaniem, że opiniowana rozprawa Pani mgr Alicji Płuciennik pt. „Algorytmy integracji danych i uczenia maszynowego w diagnostyce nowotworów" zawiera samodzielne rozwiązanie ważnego i istotnego problemu naukowego, jednocześnie spełniając

wszystkie wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej Ustawie o Tytule Naukowym i Stopniach Naukowych.

W związku z tym stawiam wniosek o dopuszczenie rozprawy doktorskiej do publicznej obrony.