dr hab. Grzegorz Marcin Wójcik, prof. UMCS
Kierownik Katedry Neuroinformatyki i Inżynierii Biomedycznej
Uniwersytet Marii Curie-Skłodowskiej Lublinie
Instytut Informatyki
ul. Akademicka 9, 20-033 Lublin
gmwojcik@live.umcs.edu.pl

Recenzja rozprawy doktorskiej mgr inż. Aleksandra Płaczka

Tytuł rozprawy: Wsparcie diagnostyki onkologicznej w oparty o dane omiczne procesie wielokryterialnego wnioskowania parametryzowanego obrazem klinicznym pacjentów o podobnym profilu

Promotor w przewodzie: dr hab. Dariusz Mrozek, prof. PS, Katedra Informatyki Stosowanej, Politechnika Śląska

Promotor pomocniczy: dr hab. n. med. Michał Jarząb, Centrum Diagnostyki i Leczenia Chorób Piersi, Narodowy Instytut Onkologii w Gliwicach

Przedłożona do oceny rozprawa p. mgr inż. Aleksandra Płaczka została zrealizowana w ramach doktoratu wdrożeniowego i dotyczy diagnostyki chorób tarczycy ukierunkowanej na wykrywanie zmian nowotworowych oraz określanie złośliwości guzków tarczycy metodami komputerowymi.

W szczególności autor proponuje metodologię, w której stosuje się trzy podejścia:

1. Zastosowanie autorskiego algorytmu przeszukiwania tekstu do analizy danych opisowych dołączanych do wyników badań cytologicznych.

2. Zastosowanie syntetycznego modelu bayesowskiego do szacowania ryzyka klinicznego na podstawie danych zebranych z Narodowego Instytutu Onkologii – Państwowego Instytutu Badawczego.

3. Zastosowanie autorskiego modelu bayesowskiego wyspecjalizowanego w prognozowaniu złośliwości guzka tarczycy na podstawie cech omicznych oraz klinicznych.

Biorąc powyższe pod uwagę, uważam, że wybór tematu dysertacji jest właściwy i uzasadniony.

Praca została napisana w języku polskim, zawiera 162 strony. Główna jej część składa się z sześciu rozdziałów (w tym rozdział wsrępny i podsumowujący). Pracę opatrzono również streszczeniem, spisem treści i sekcją podziękowań na początku, oraz spisem tabel i rysunków na końcu. Bibliografia obejmuje 160 dobrze dobranych i ponumerowanych pozycji.

Rozdział 1 rozprawy to rozdział wstępny. Zawarte w nim są motywacje stanowiące uzasadnienie do przeprowadzenia badań, sformułowane cele rozprawy i hipotezy badawcze, kontekst badań oraz podstawowe ograniczenia.

Sformułowane przez autora główne cele rozprawy stanowią:

1. „Opracowanie modelu wyznaczającego syntetyczne ryzyko kliniczne złośliwości guzka tarczycy, wykorzystującego surowe dane kliniczne gromadzone w szpitalnych bazach danych."

2. „Wykorzystanie tego ryzyka do zwiększenia współczynnika wpływu czynników klinicznych na ostateczny wynik klasyfikacji modeli predykcyjnych, bazujących na kombinacji danych klinicznych i omicznych."

Natomiast główna hipoteza badawcza brzmi następująco:

- „Stosowanie w modelach wspierających diagnostykę onkologiczną guzka tarczycy, syntetycznego ryzyka klinicznego, opartego o retrospektywną analizę zanonimizowanych danych klinicznych dużej grupy chorych, na równi z danymi molekularnymi pozwala na zmniejszenie rozbieżności in-

terpretacyjnych klinicysty między wynikiem badania molekularnego i obrazem klinicznym."

W dalszych rozdziałach autor dąży do realizacji celów oraz weryfikacji hipotezy badawczej wraz z weryfikacją sformułowanymi na wstępie hipotezami pomocniczymi.

Najpierw jednak mamy do czynienia z dwoma rozdziałami wprowadzającymi czytelników w problematykę diagnozowania guzków tarczycy oraz statystyk bayesowskich.

I tak w Rozdziale 2 autor opisuje przedkładając dokładnie opracowany schemat postępowania w procesie diagnostycznym drogę jaką przechodzi pacjent od pierwszej konsultacji lekarskiej do postawienia diagnozy, zaplanowania leczenia oraz leczenia. Opis jest przejrzysty i zawiera wszelkie możliwe ścieżki przebiegu takiego procesu. Następnie prezentowane są rodzaje dostępnej informacji, która może dalej być wykorzystana w procesie diagnozy oraz jej wsparcia. Są to profile kliniczny i transkryptomiczny pacjenta, które wywołują do dalszych rozważań zagadnienie jakości informacji odgrywającej kluczową rolę w algorytmach uczenia maszynowego opisywanych przez autora w sekcji 2.3. To uczeniu maszynowemu autor poświęca najwięcej miejsca w Rozdziale 2. Otrzymujemy dobry opis wad i zalet podstawowych klasyfikatorów wykorzystywanych w Machine Learningu wraz z wymienieniem ich zalet oraz wad. Na szczególne podkreślenie zasługuje osadzenie przez autora rozważań w istniejącym stanie wiedzy, podparty licznymi odwołaniami do najnowszych prac naukowych w interesującym go obszarze.

Rozdział 3 zawiera opis statystyki bayesowskiej. Jak wiemy chociażby z lektury streszczenia pracy, to sieci bayesowskie odegrały kluczową rolę w zaprojektowanych przez autora modelach. Umieszczenie Rozdziału 3 wraz zaawansowanych aparatem matematycznym opisującym teorię bayesowską oraz jej za-

stosowanie i znaczenie dla algorytmów wnioskowania jest zatem w pełni uzasadnione. Po lekturze rozdziałów 2 i 3 nawet nieobeznany z tematyką badań czytelnik wprowadzony zostaje w zagadnienie tak od strony medycznej jak i matematyczno-informatycznej co przy realizacji badań interdyscyplinarnych ma istotne znaczenie.

W Rozdziale 4 zamieszczony został opis materiałów i metod wykorzystywanych i stosowanych w ramach realizacji rozprawy. I tak otrzymujemy szczegółowe opisy danych klinicznych oraz mikromacierzowych. Następnie autor przedstawia z uzasadnieniem wybór odpowiedniej sieci bayesowskiej. Jak zazwyczaj w uczeniu maszynowym – dobre opracowanie wstępne danych może stanowić wręcz o powodzeniu całego przedsięwzięcia. Autor przedstawia w sposób satysfakcjonujący zarówno metodę dyskretyzacji próbek wraz z miarami oceny tych próbek jak i zbiór uczący i walidacyjny wykorzystywany w modelu.

Rozdział 5 zawiera opis wyników badań prowadzonych przez autora. Pod względem objętości jest to rozdział największy i w mojej ocenie najważniejszy. Zawiera 47 stron, 26 rysunków, 4 tabele.

W tym rozdziale autor krok po kroku realizuje postawione na wstępie cele, pozytywnie weryfikuje też hipotezę badawczą wraz z obiema hipotezami pomocniczymi. To oznacza, że:

- Stosowanie w modelach wspierających diagnostykę onkologiczną guzka tarczycy, syntetycznego ryzyka klinicznego, opartego o retrospektywną analizę zanonimizowanych danych klinicznych dużej grupy chorych, na równi z danymi molekularnymi pozwala na zmniejszenie rozbieżności interpretacyjnych klinicysty między wynikiem badania molekularnego i obrazem klinicznym.

- Wartość predykcyjną zestawu genów, stosowanego do różnicowania złośliwości w guzkach tarczycy, można zwiększyć, wykorzystując syntetyczną

zmienną reprezentującą kliniczne ryzyko złośliwości zmiany, bez ujemnego wpływu słabszej jakościowo cechy klinicznej na siłę predykcyjną połączonego zestawu cech.

- Wykorzystanie syntetycznej zmiennej klinicznej, zamiast surowych danych klinicznych, pozwoli na wyeliminowanie wpływu subiektywizmu w ocenie materiału biopsyjnego na jakość klasyfikacji.

Nie jest celem recenzji pisanie streszczenia rozdziału 5 recenzowanej rozprawy doktorskiej. Jakkolwiek na szczególnie pozytywną ocenę zasługuje wybór metodologii zastosowanej przez p. mgr. inż. Aleksandra Płaczka, w wyniku czego w mojej ocenie powstał innowacyjny protokół oraz tzw. Data Science Pipeline do wykorzystania w wsparciu procesu diagnostycznego osób z chorobami tarczycy.

Algorytmy zaproponowane przez autora są nowe, autorskie, według mojej wiedzy nie zostały wykorzystane wcześniej, stanowią istotny wkład w dyscyplinę.

Na szczególne podkreślenie zasługuje liczba badań i testów przeprowadzonych przez autora. należy zauważyć, że badania prowadzone w ramach przewodu doktorskiego zbiegły się w czasie z pandemią COVID-19. Nawet bez pandemii, pozyskanie danych medycznych z archiwów szpitalnych bywa niezwykle trudne. Tu jednak mamy do czynienia z profesjonalnym, systematycznym podejściem do zagadnienia.

W Rozdziale 6 autor w ramach podsumowania gromadzi w jednym miejscu wnioski wynikające z opisanych w Rozdziale 5 badań, zamieszcza dyskusję, podsumowania oraz wyznacza potencjalne kierunki rozwoju badań w przyszłości. Na Rys. 6.1 autor przedstawia zresztą diagram wdrożeniowy, który może znaleźć znaleźć w/w zastosowanie.

Pytania i uwagi krytyczne

Praca jest napisana poprawnie językowo i stylistycznie, bardzo dobrze złożona (LaTeX). Nie znajduję w niej błędów i niedociągnięć technicznych, które miałyby wpływ na jej ogólny odbiór i merytoryczne przesłanie.

Podczas lektury pojawiło się u mnie natomiast kilka pytań, do których chciałbym by doktorant ustosunkował się w przypadku dopuszczenia do obrony:

- Jaka jest wydajność czasowa algorytmu RetroNGSC przedstawionego w sekcji 4.1? Przy tysiącach pacjentów i dziesiątkach tysięcy nakłuć może to mieć istotne znaczenie?
- Jak ta wydajność ma się do wydajności klasyfikacji przy pomocy algorytmu C4.5 opisanego w sekcji 5.1, o którym mowa na stronie 85 rozprawy?
- Czy istnieje różnica w wydajności (czasowej) tworzenia lub użycia modeli prognostycznych integrujących dane kliniczne i omiczne, przedstawionych w sekcji 4.8? Czy są one porównywalne? Czy istnieją wyraźne różnice?
- Ciekawi mnie samego czy można byłoby algorytm RetroNGSC zaimplementować na GPU?

Wszystkie powyższe pytania mają jedynie charakter pytań z natury dociekliwych i nie mają wpływu na moje jednoznacznie pozytywne wrażenie z lektury pracy doktorskiej pana Aleksandra Płaczka.

Rekomendacja

Recenzowany przeze mnie doktorat ma charakter wdrożeniowy. Oprócz wiodącego wsparcia promotora w osobie dr hab. Dariusza Mrozka, profesora Politechniki Śląskiej ze strony medycznej promotorstwo pomocnicze sprawował dr. hab.

med. Michał Jarząb z Narodowego Instytutu Onkologii im. Marii Skłodowskiej-Curie – Państwowego Instytutu Badawczego (NIO-PIB Oddział w Gliwicach). Na podkreślenie zasługuje także współpraca z przemysłem, w tym wypadku z firmami WASKO S.A. oraz GABOS Software Sp. z o. o. Tego typu doktoraty są potrzebne i mają istotne znaczenie dla gospodarki.

Badania prowadzone w ramach realizacji projektu, referowane w rozprawie opublikowano częściowo w wysokopunktowanym czasopiśmie:

- Płaczek, A., Płuciennik, A., Kotecka-Blicharz, A., Jarzab, M., & Mrozek, D. (2020). Bayesian assessment of diagnostic strategy for a thyroid nodule involving a combination of clinical synthetic features and molecular data. IEEE Access, 8, 175125-175139.

oraz prezentowano na konferencjach naukowych o zasięgu międzynarodowym:

- Płaczek, A., Płuciennik, A., Pach, M., Jarząb, M., & Mrozek, D. (2019, May). The role of feature selection in text mining in the process of discovering missing clinical annotations: Case study. In International Conference: Beyond Databases, Architectures and Structures (pp. 248-262). Springer, Cham.

- Psiuk-Maksymowicz, K., Jaksik, R., Płaczek, A., Gruca, A., Student, S., Borys, D., ... & Świerniak, A. (2019, September). BioTest-Remote Platform for Hypothesis Testing and Analysis of Biomedical Data. In Polish Conference on Biocybernetics and Biomedical Engineering (pp. 152-165). Springer, Cham.

Moja ocena rozprawy doktorskiej p. mgr. inż. Aleksandra Płaczka jest jednoznacznie pozytywna.

Uważam, że rozprawa doktorska mgr. inż. Aleksandra Płaczka spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018

r. o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668), dlatego zwracam się do Wysokiej Rady Dyscypliny Informatyki Technicznej i Telekomunikacji Politechniki Śląskiej o dopuszczenie mgr. inż. Aleksandra Płaczka do dalszych etapów przewodu doktorskiego.

Dodatkowo ze względu na ilość przeprowadzonych badań, dorobek publikacyjny i znaczenie wyników, wnoszę o wyróżnienie rozprawy.

Kierownik Katedry
dr hab. Grzegorz M. Wójcik
prof. UMCS

000001353
Uniwersytet Marii Curie-Skłodowskiej
Wydział Matematyki, Fizyki i Informatyki
Katedra Neuroinformatyki i Inżynierii Biomedycznej
ul. Akademicka 9
20-033 Lublin, tel. 81 537-29-40

Lublin, 2022-08-31