dr hab. Zuzanna Szymańska
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego,
Uniwersytet Warszawski,
zk.szymanska@uw.edu.pl

Recenzja Pracy Doktorskiej

Autor: mgr inż. Agata Wilk
Tytuł Pracy: „Opracowanie nowych algorytmów uczenia maszynowego dla heterogenicznych danych biomedycznych"
Promotor: prof. dr hab. inż. Krzysztof Fujarewicz

Recenzję rozprawy doktorskiej sporządzono na prośbę Przewodniczącej Rady Dyscypliny Inżynierii Biomedycznej Politechniki Śląskiej Prof. dr hab. inż. Ewy Piętki, wyrażoną w piśmie RDIB.0211.33.2024 z dnia 20 czerwca 2024 roku.

Charakterystyka Rozprawy

Rozprawa doktorska „Opracowanie nowych algorytmów uczenia maszynowego dla heterogenicznych danych biomedycznych" została przygotowana przez panią mgr inż. Agatę Wilk w Katedrze Inżynierii i Biologii Systemów Politechniki Śląskiej. Autorka rozprawy wyznaczyła cztery cele badawcze: i) opisanie heterogeniczności występującej na różnych poziomach w danych biomedycznych, ii) opracowanie algorytmów do analizy i modelowania heterogenicznych populacji oraz kohort, iii) przedstawienie zagadnień związanych ze strukturalną heterogenicznością danych, oraz iv) zaproponowanie metod agregacji, które umożliwiają wykorzystanie informacji z różnej liczby wektorów cech odpowiadających poszczególnym modelowanym obiektom. Ponadto autorka sformułowała dwie tezy: i) indywidualizacja modeli dla kohorty obiektów zmniejsza ryzyko złego uwarunkowania numerycznego zadania estymacji parametrów, oraz ii) zastosowanie agregacji przy różnej liczbie wektorów cech dla poszczególnych obiektów prowadzi do poprawy jakości predykcji modelu w porównaniu do użycia tylko jednego wektora na obiekt.

Rozprawę doktorską mgr inż Agaty Wilk stanowi cykl siedmiu artykułów:

1. Izabela Zarczynska, Monika Gorska-Arcisz, Alexander Jorge Cortez, Katarzyna Aleksandra Kujawa, **Agata Małgorzata Wilk**, Andrzej Cezary Skladanowski, Aleksandra Stanczak, Monika Skupinska, Maciej Wieczorek, Katarzyna Marta Lisowska, Rafal Sadej i Kamila Kitowska. „p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer". Cells 10(12):3363 (2021).

2. Katarzyna Mrowiec, Julia Debik, Karol Jelonek, Agata Kurczyk, Lucyna Ponge, **Agata Wilk**, Marcela Krzempek, Guro F. Giskeødegård, Tone F. Bathen i Piotr Widłak. „Profiling of serum metabolome of breast cancer: multicancer features discriminate between healthy women and patients with breast cancer". Frontiers in Oncology 14 (2024).

3. **Agata Małgorzata Wilk**, Krzysztof Łakomiec, Krzysztof Psiuk-Maksymowicz i Krzysztof Fujarewicz. „Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling". Scientific Reports 12.1 (2022).

4. Agata Kurczyk, Marta Gawin, Mykola Chekan, **Agata Wilk**, Krzysztof Łakomiec, Grzegorz Mrukwa, Katarzyna Frątczak, Joanna Polanska, Krzysztof Fujarewicz, Monika Pietrowska i Piotr Widlak. „Classification of Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach". International Journal of Molecular Sciences 21.17 (2020).

5. **Agata Małgorzata Wilk**, Emilia Kozłowska, Damian Borys, Andrea D'Amico, Krzysztof Fujarewicz, Izabela Gorczewska, Iwona Debosz-Suwinska, Rafał Suwinski, Jarosław Smieja i Andrzej Swierniak. „Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer". Translational Lung Cancer Research 12.7 (2023), s. 1372–1383.

6. **Agata Małgorzata Wilk**, Emilia Kozłowska, Damian Borys, Andrea D'Amico, Izabela Gorczewska, Iwona Debosz-Suwińska, Seweryn Gałecki, Krzysztof Fujarewicz, Rafał Suwiński i Andrzej Świerniak. „Improving the Predictive Ability of Radiomics-Based Regression Survival Models Through Incorporating Multiple Regions of Interest". The Latest Developments and Challenges in Biomedical Engineering. Red. Paweł Strumiłło, Artur Klepaczko, Michał Strzelecki i Dorota Bociąga. Cham: Springer Nature Switzerland, 2024, s. 163–173.

7. **Agata Małgorzata Wilk**, Andrzej Swierniak, Andrea d'Amico, Rafał Suwiński, Krzysztof Fujarewicz i Damian Borys. „Towards the use of multiple ROIs for radiomicsbased survival modelling: finding a strategy of aggregating lesions". arXiv: 2405.17668, (2024).

Pierwszym celem badawczym, który postawiła sobie Autorka, było opisanie heterogeniczności występującej na różnych poziomach w danych biomedycznych. Odnosząc się do pracy Zarczynska et al. (2021), Autorka omówiła heterogeniczność na poziomie komórkowym oraz jej wpływ na efektywną analizę danych [1]. Na podstawie przeprowadzonych analiz danych genomicznych i transkryptomicznych wykazała, że w przypadku małych prób zmienność wynikająca z różnic pomiędzy typami komórek może maskować poszukiwany mechanizm biologiczny. W związku z tym bardziej efektywnym podejściem może być analiza danych dla poszczególnych typów komórek niezależnie, a następnie identyfikacja wspólnych elementów. Praca jest bardzo interesująca, stanowi wzorcowe połączenie badań eksperymentalnych z teoretyczną analizą bioinformatyczną, a uzyskane wyniki mają potencjalnie istotne

znaczenie terapeutyczne. Warto zaznaczyć, że Autorka rozprawy była jedną z dwóch osób odpowiedzialnych za opracowanie części bioinformatycznej. W nawiązaniu do kolejnej pracy zawartej w rozprawie, Autorka rozprawy omawia heterogeniczność na poziomie tkankowym, przedstawiając metabolomiczną charakterystykę raka piersi oraz wyniki badań dotyczących identyfikacji wspólnej sygnatury dla różnych typów nowotworów [2]. Praca ta dostarcza wyników o potencjalnie istotnym znaczeniu diagnostycznym. Autorka rozprawy wskazuje, że pomimo różnic między typami nowotworów, odpowiednie przetwarzanie tak heterogenicznych danych pozwala na wyselekcjonowanie wspólnej sygnatury, umożliwiającej precyzyjną klasyfikację pacjentek z rakiem piersi, przynajmniej w badanych populacjach polskiej i norweskiej. Dr inż. Agata Wilk była również odpowiedzialna za kluczową część bioinformatyczną tego badania. Kolejnym omawianym przykładem heterogeniczności danych biomedycznych, tym razem na poziomie populacji, są dane epidemiczne dotyczące dynamiki COVID-19. W pracy Wilk et al. (2022) zaproponowano algorytm indywidualizowanej estymacji parametrów modelu, co pozwoliło na uzyskanie niższych błędów predykcji niż w przypadku niezależnej lub wspólnej estymacji modeli epidemiologicznych dla poszczególnych krajów europejskich [3]. Uzasadniając potrzebę opracowania nowej metody szacowania parametrów, Autorka opisuje specyfikę danych populacyjnych, w tym przypadku dotyczących zachorowań na COVID-19. Na koniec, niejako domykając pierwszy cel badawczy, Autorka rozprawy analizuje heterogeniczność danych pochodzących z obrazowania medycznego. Punktem odniesienia jest tutaj piąta praca wchodząca w skład cyklu, dotycząca przewidywania ryzyka przerzutów odległych dla pacjentów z niedrobnokomórkowym rakiem płuc na podstawie danych klinicznych i radiomicznych [5].

Drugim celem badawczym rozprawy było opracowanie algorytmów do analizy i modelowania heterogenicznych populacji oraz kohort. Cel ten został osiągnięty przez Autorkę dzięki jej kluczowemu wkładowi w zaproponowanie nowego podejścia do szacowania parametrów modeli epidemiologicznych, co dostarczyło nowe narzędzia do planowania strategii kontroli pandemii [3]. W pracy wykazano, że zastosowanie indywidualizowanego podejścia do estymacji parametrów modeli dla kohorty obiektów pozwala na ograniczenie ryzyka niewłaściwego uwarunkowania numerycznego, jednocześnie zachowując indywidualny charakter obiektów. Moim zdaniem zaproponowany algorytm estymacji parametrów na podstawie danych epidemicznych jest wyjątkowy, a wyniki pracy, kwantyfikujące skuteczność poszczególnych administracyjnych metod ograniczania rozprzestrzeniania pandemii COVID-19, są niezwykle interesujące. Warto podkreślić, że praca została opublikowana w renomowanym czasopiśmie, co dodatkowo świadczy o jej wysokiej wartości naukowej.

Trzecim celem badawczym rozprawy było przedstawienie zagadnień związanych ze strukturalną heterogenicznością danych, ze szczególnym uwzględnieniem sytuacji, w której liczba dostępnych wektorów cech różni się dla poszczególnych obiektów. Autorka rozważa przypadki przestrzennego badania molekularnego zdrowych oraz zmienionych nowotworowo tkanek tarczycy w celu poprawy diagnostyki [4], a także analizy danych obrazowych w kontekście przewidywania ryzyka przerzutów w regionalnie zaawansowanym nowotworze płuc [5, 6, 7]. Autorka rozprawy w sposób

bardzo klarowny przedstawia zagadnienie strukturalnej heterogeniczności w omawianych przypadkach, wyjaśniając różnice między wspomnianymi typami danych. I tak, w technikach molekularnych każdy pomiar dotyczy określonego punktu siatki, którą „pokryta" jest tkanka, co skutkuje różną liczbą wektorów cech dla poszczególnych próbek. Natomiast w obrazowaniu medycznym wektory cech mogą dodatkowo różnić się istotnością. Na przykład jedną z kluczowych cech różnicujących wektory cech jest wielkość (objętość) regionu zainteresowania, z jakiego zostały wyznaczone.

Czwartym celem badawczym sformułowanym w rozprawie było zaproponowanie metod agregacji wektorów cech lub wyników modeli, które mogą być stosowane zarówno w modelach klasyfikacyjnych [4], jak i przeżyciowych [6, 7]. Zaproponowane przez Autorkę metody agregacji są szczególnie przydatne w przypadku, gdy dane wejściowe do budowanych modeli predykcyjnych cechują się heterogenicznością strukturalną, na przykład wtedy, gdy dla niektórych obiektów dostępnych jest więcej niż jeden wektor cech, a ich liczba jest zmienna. W pracy Kurczyk et al. (2022) dla problemu klasyfikacji raka tarczycy z wykorzystaniem danych z obrazowania proteomicznego zaproponowano podejście oparte na agregacji wyników predykcji dla pojedynczych widm, co umożliwiło osiągnięcie znacznie wyższej jakości klasyfikacji niż w przypadku użycia jedynie uśrednionego widma [4].W mojej opinii bardzo interesujący jest cykl prac poświęconych przewidywaniu ryzyka przerzutów odległych u pacjentów z niedrobnokomórkowym rakiem płuc, oparty na analizie danych klinicznych i radiomicznych [5,6,7]. W tych publikacjach Autorka rozprawy stopniowo rozwija stosowane metody, co prowadzi do uzyskiwania coraz lepszych wyników predykcji. Na podstawie osiągniętych przez Doktorantkę rezultatów można stwierdzić, że założony cel badawczy został w pełni zrealizowany. Zaproponowane metody agregacji wektorów cech lub wyników modeli umożliwiają uwzględnienie wielu ognisk nowotworu oraz ich zróżnicowania, co znacząco poprawia jakość predykcji w różnych modelach.

Reasumując, zastosowanie zaproponowanych strategii, takich jak indywidualizacja estymacji parametrów czy agregacja wyników modelowania, pozwoliło osiągnąć znacznie lepsze rezultaty w porównaniu do metod opartych na maksymalnej generalizacji i uśrednianiu. Sformułowane przez Doktorantkę dwie tezy:

1. Indywidualizacja modeli dla kohorty obiektów pozwala zmniejszyć ryzyko złego uwarunkowania numerycznego zadania estymacji parametrów.

2. Zastosowanie agregacji w przypadku różnej liczby wektorów cech dla poszczególnych obiektów skutkuje poprawą jakości predykcji modelu względem wykorzystania tylko jednego wektora na obiekt.

zostały więc w satysfakcjonujący sposób poparte rezultatami opublikowanymi w przedłożonych pracach i omówione w rozprawie.

Ocena Merytoryczna

Przygotowaną przez mgr inż. Agatę Wilk rozprawę pt. „Opracowanie nowych algorytmów uczenia maszynowego dla heterogenicznych danych biomedycznych" oceniam bardzo wysoko zarówno pod względem wyboru tematu, jak i zaproponowanych rozwiązań. Doktorantka zrealizowała ambitne cele badawcze, wykazując się przy tym dużą pomysłowością i starannością. Opublikowane przez nią wyniki badań świadczą o jej szerokiej wiedzy z zakresu nauk inżynieryjno-technicznych, w szczególności w dziedzinie inżynierii biomedycznej.

Chciałabym mocno podkreślić osiągnięcia Doktorantki, które zostały opublikowane w pracach, gdzie występuje jako pierwszy autor [3,5,6,7]. W szczególności zwracam tu uwagę na publikację, w której zaproponowano ciekawą metodę szacowania parametrów modeli epidemiologicznych [3]. Dzięki tej pracy możliwe było opracowanie wiarygodnych narzędzi wspierających planowanie strategii kontroli pandemii. W mojej opinii, jednym z głównych wyzwań związanych z praktycznym zastosowaniem modeli matematycznych jest brak ich odpowiedniej kalibracji. Z tego powodu z dużym uznaniem odnoszę się do rozwiązania zaproponowanego przez Doktorantkę, które opiera się na indywidualizowanym podejściu do estymacji parametrów dla kohorty obiektów. Uważam to podejście za pomysłowe i nowatorskie, ponieważ pozwala ono na precyzyjniejsze dopasowanie modeli do zróżnicowanych danych populacyjnych, co jest kluczowe w przypadku tak dynamicznych zjawisk jak pandemie.

Uwagi i Sugerowane Poprawki

Praca została napisana bardzo starannie i przystępnie, a jej poszczególne części tworzą spójną całość, co umożliwia zrozumienie idei prowadzonych badań oraz ocenę zaproponowanych rozwiązań algorytmicznych. Uważam, że nawet osoby niezwiązane bezpośrednio z omawianą tematyką mogą łatwo przyswoić sobie poszczególne zagadnienia i poszerzyć swoją wiedzę dzięki lekturze rozprawy. Mimo starannej lektury, dostrzegłam tylko jeden drobny błąd edycyjny (na stronach 16 i w angielskiej wersji na stronie 51 pojawia się urwane zdanie). Drugą, równie drobną uwagą krytyczną jest niewielka czytelność napisów na niektórych obrazkach.

Podsumowanie

Dorobek naukowy mgr inż. Agaty Wilk obejmuje łącznie 25 publikacji, z czego 17 artykułów jest indeksowanych w Web of Science (18. artykuł stanowi korektę innej pracy). H-index Autorki według bazy Web of Science wynosi 4, a jej prace zostały cytowane łącznie 40 razy. Natomiast według Google Scholar H-index Autorki rozprawy również wynosi 4, przy 67 cytowaniach. Uważam, że na tym etapie kariery naukowej jest to znaczące osiągnięcie.

Podsumowując, stwierdzam, że recenzowana praca spełnia wymagania określone dla rozpraw doktorskich przez Ustawę z dnia 20 lipca 2018 roku - Prawo o szkolnictwie wyższym i nauce (j.t. Dz. U. z 2023 r. poz. 742, z późn. zm.). Biorąc pod

uwagę osiągnięte wyniki, dorobek naukowy Doktorantki oraz obowiązujące przepisy dotyczące stopni i tytułów naukowych, rekomenduję Radzie Dyscypliny Inżynierii Biomedycznej Politechniki Śląskiej dopuszczenie pracy do kolejnych etapów postępowania doktorskiego. Jednocześnie, biorąc pod uwagę interdyscyplinarność badań, wysoką jakość uzyskanych wyników oraz znaczący dorobek naukowy Doktorantki, wnioskuję o wyróżnienie rozprawy doktorskiej mgr inż. Agaty Wilk.

Zuzanna Szymańska

dr hab. Zuzanna Szymańska

Warszawa, 4 września 2024