

Dr hab. inż. Paweł Piotr Łabaj, prof. UJ
Małopolskie Centrum Biotechnologii
Uniwersytet Jagielloński
Gronostajowa 7a, 30-387 Kraków
pawel.labaj@uj.edu.pl

Kraków, 12.05.2025

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Informatyczna analiza danych multiomicznych w oparciu o techniki uczenia maszynowego i metody statystyczne
Autor rozprawy:	mgr inż. Jagoda Głowacka-Walas
Promotor rozprawy:	dr hab. inż. Tomasz Gambin, prof. uczelni
Dziedzina:	nauki inżynieryjno-techniczne
Dyscyplina:	Informatyka Techniczna i Telekomunikacja

Rozwój wysokoprzepustowych biomedycznych technik pomiarowych pozwolił w ostatnich latach na znaczny postęp w naukach biologicznych i medycznych. Umożliwił również na dokładniejsze przyjrzenie się wielu chorobom ale tym z punktu widzenia szerokiej/głębokiej charakteryzacji molekularnej. Doskonałym przykładem wykorzystania tych możliwości jest rozprawa doktorska mgr inż. Jagody Głowackiej-Walas zatytułowana "Informatyczna analiza danych multiomicznych w oparciu o techniki uczenia maszynowego i metody statystyczne". Praca ta wpisuje się w aktualne trendy badawcze, koncentrując się na opracowaniu narzędzi i metod wspomagających diagnostykę oraz identyfikację biomarkerów chorobowych poprzez analizę danych multiomicznych, ze szczególnym uwzględnieniem chorób rzadkich. Autorka, motywowana wyzwaniami zidentyfikowanymi w projektach klinicznych EPISTOP i EPIMARKER, podjęła się opracowania modeli predykcyjnych dotyczących występowania napadów padaczkowych lub lekooporności u dzieci ze stwardnieniem guzowatym.

Należy podkreślić, że integracja danych multiomicznych, będąca sednem prezentowanej rozprawy, jest zadaniem niezwykle złożonym i obciążonym licznymi wyzwaniami. Już na etapie przetwarzania wstępnego wyzwaniem jest ogrom i złożoność danych, gdzie pojedyncze pomiary z wykorzystaniem technik szerokoprzepustowych (np. RNA-Seq) generują pliki o znacznych rozmiarach, wymagające odpowiedniej mocy obliczeniowej i skrupulatności w wieloetapowej obróbce. Sama integracja różnorodnych warstw danych omicznych stawia przed badaczami konieczność wyboru odpowiedniej strategii – od wczesnej integracji, polegającej na konkatenacji danych, która jednak może prowadzić do problemu wysokiej wymiarowości i nierównomiernego wpływu poszczególnych typów danych na model, po bardziej zaawansowane metody, takie jak integracja pośrednia, która

choć potencjalnie odkrywa głębsze zależności biologiczne, jest kosztowna obliczeniowo, a jej wyniki bywają trudne do bezpośredniej interpretacji biologicznej. Rozprawa doktorska porusza również problem małej liczby obserwacji w stosunku do liczby cech (tzw. problem $p \gg n$), co jest szczególnie dotkliwe w badaniach nad chorobami rzadkimi i zwiększa ryzyko nadmiernego dopasowania modeli. Dodatkowymi komplikacjami są częste braki danych, nierzadko o charakterze blokowym, oraz niezbalansowanie klas w badanych kohortach.

Motywnym przewodnim rozprawy jest doświadczenie nabywane przez Autorkę wraz z postępowaniem prac nad projektami, najpierw EPISTOP, a później EPIMARKER.

Pierwszy z projektów (EPISTOP) był prowadzony w latach 2013-2019 i skupiał się na obserwacji 101 dzieci ze stwardnieniem guzowatym (TSC) od urodzenia do drugiego roku życia. Stwardnienie guzowate to rzadka choroba genetyczna, w której padaczka jest jednym z najczęstszych i najpoważniejszych objawów, rozwijającym się często w pierwszym roku życia i prowadzącym do znaczących powikłań neurorozwojowych. Główne założenia i cele projektu EPISTOP można podzielić na dwa kluczowe obszary: **i) cel kliniczny:** Ocena skuteczności wczesnego, prewencyjnego leczenia przeciwpadaczkowego u niemowląt z TSC, jeszcze przed klinicznym ujawnieniem się napadów. Założeniem było, że takie postępowanie może zatrzymać rozwój padaczki lub znacząco złagodzić jej przebieg, co w konsekwencji miało prowadzić do poprawy wyników leczenia i jakości życia pacjentów. Wyniki tej części badania wykazały, że leczenie prewencyjne faktycznie zmniejsza ryzyko rozwoju padaczki, opóźnia jej pojawienie się oraz łagodzi ciężkość i ryzyko lekooporności; **ii) cel molekularny:** Identyfikacja biomarkerów molekularnych, które pozwoliłyby na przewidywanie ryzyka rozwoju padaczki u dzieci z TSC. Odkrycie takich biomarkerów miało umożliwić wczesne wdrożenie terapii, jeszcze przed pojawieniem się objawów klinicznych, a także przyczynić się do lepszego zrozumienia mechanizmów leżących u podstaw epileptogenezy w TSC. W ramach tego celu planowano: i) Analizę danych dla pojedynczych warstw "omicznych" w celu identyfikacji specyficznych biomarkerów; ii) Integrację danych multiomicznych w celu opracowania modeli predykcyjnych pozwalających przewidzieć wystąpienie napadów padaczkowych. Analiza ta obejmowała dane transkryptomiczne, proteomiczne, metaboliczne, miRNA oraz dane z sekwencjonowania całego genomu (WGS). Projekt zakładał podłużną obserwację pacjentów i pobieranie próbek biologicznych (głównie krwi) w kilku punktach czasowych, aby móc uchwycić zmiany molekularne poprzedzające rozwój padaczki. Do analiz predykcyjnych wykorzystywano próbki pobrane w najwcześniejszym możliwym momencie, przed wystąpieniem napadów.

Drugim z projektów na którym bazuje rozprawa to projekt EPIMARKER, realizowany w latach 2017-2022, stanowił rozwinięcie badań zainicjowanych w projekcie EPISTOP, koncentrując się na dalszym zgłębianiu problematyki padaczki u pacjentów ze stwardnieniem guzowatym (TSC). Główne założenia i cele projektu EPIMARKER obejmowały: **i) identyfikację biomarkerów lekooporności:** Kluczowym celem było znalezienie biomarkerów, które pozwoliłyby na wczesne rozpoznanie ryzyka rozwoju padaczki lekoopornej u niemowląt i małych dzieci cierpiących na padaczkę związaną z

TSC; ii) **identyfikację biomarkerów nawrotu padaczki**: Drugim istotnym celem była identyfikacja biomarkerów umożliwiających przewidywanie nawrotu napadów padaczkowych po odstawieniu leczenia przeciwpadaczkowego u dzieci i nastolatków z TSC, u których wcześniej udało się osiągnąć remisję. Należy jednak zaznaczyć, że z powodu trudności w rekrutacji pacjentów i ograniczonej dostępności danych, wyniki prezentowane w rozprawie doktorskiej koncentrują się głównie na pierwszym obszarze, czyli lekooporności. W ramach części multiomicznej projektu, która jest szczególnie istotna dla kontekstu rozprawy, postawiono następujące zadania: i) **Opracowanie algorytmu klasyfikacyjnego dla ryzyka lekooporności**: Dążono do stworzenia algorytmu, który na podstawie zintegrowanych danych klinicznych, wyników badań elektroencefalograficznych (EEG) oraz biomarkerów molekularnych, byłby w stanie ocenić ryzyko rozwoju padaczki lekoopornej u pacjentów z TSC; ii) **Identyfikacja biomarkerów predykcyjnych dla nawrotu padaczki**: Poszukiwano biomarkerów, które mogłyby wskazywać na prawdopodobieństwo powrotu napadów po zaprzestaniu farmakoterapii (ten cel, jak wspomniano, napotkał ograniczenia w realizacji z powodu trudności rekrutacyjnych). Projekt EPIMARKER miał zatem na celu dostarczenie narzędzi i wiedzy, które mogłyby przyczynić się do bardziej spersonalizowanego i skutecznego leczenia padaczki w tej specyficznej grupie pacjentów.

Wkład Autorki w badania będące podstawą doktoratu, realizowane w ramach projektów EPISTOP i EPIMARKER, był znaczący. W projekcie EPISTOP Doktorantka uczestniczyła zarówno w analizie danych dla pojedynczych "omik", jak i w analizach multiomicznych. Jej wkład zaowocował m.in. współdzielonym pierwszym autorstwem publikacji w *Nature Communications*. W zakresie analiz pojedynczych "omik", choć metodologia była opracowywana wspólnie z Kamilem Sijko (TTSI) oraz zespołem molekularnym projektu EPISTOP, to implementacja analiz, ich wykonanie, przygotowanie wyników oraz wizualizacji do publikacji stanowiły całkowity wkład Autorki, co dodatkowo potwierdza udostępniony przez nią kod na platformie GitHub. W części multiomicznej projektu EPISTOP metodologia również była pracą zespołową, jednak implementacja metodologii wykorzystującej algorytmy uczenia maszynowego oraz wykonanie analiz zostało przeprowadzone przez Autorkę oraz Kamila Sijko. Autorka zaproponowała także dodatkowe uzupełnienia, jak analiza permutacyjna, i przygotowała wyniki do publikacji. Podobnie w badaniu EPIMARKER, metodologia stanowiła pracę zespołową, a jej implementacja została wykonana przez Autorkę oraz Kamila Sijko, natomiast przygotowanie danych oraz wykonanie analiz multiomicznych stanowiło wkład Doktorantki, która pełniła również funkcję kierownika projektu ze strony TTSI.

Głównym wkładem Autorki w niniejszą rozprawę jest opracowanie potoku przetwarzania danych multiomicznych nazwanego playOmics, opisanego w rozdziale szóstym. Chociaż domyślna metodologia wykonywania analiz była efektem pracy zespołowej w TTSI, a Laura Bąkała (TTSI) zaproponowała zmianę metody logowania artefaktów, to pozostałe prace, w tym przełożenie metodologii na pakiet w języku R, rozszerzenie metodologii oraz zakresu możliwych do wykonania analiz, implementacja interfejsu graficznego, implementacja dodatkowych narzędzi (m.in. w obszarze selekcji cech, uczenia

zespołowego, wyjaśnialności predykcji, testów permutacyjnych) oraz dokumentacja, stanowią wkład Autorki. Kod źródłowy pakietu oraz obraz dockerowy zostały udostępnione publicznie.

Autorka wniosła również znaczący wkład w implementację potoku playOmics w projekcie DIPGen w ramach systemu SAMGEN, gdzie przygotowanie danych do analizy i integracja źródeł danych stanowiły pracę zespołową (TTSI) ze znacznym udziałem Doktorantki, podobnie jak implementacja funkcji playOmics w systemie GeneSense.

Układ rozprawy oddaje dość złożoną materię pracy związaną z jej wdrożeniowym charakterem. W związku z tym mamy tutaj aż osiem rozdziałów, które systematycznie prowadzą czytelnika przez poszczególne projekty oraz rozwój metodologii związany z kolejnymi etapami badań i analiz:

- **Rozdział Pierwszy (Wprowadzenie):** Wprowadza w tematykę badań multiomicznych, przedstawia cel i zakres pracy, tezy badawcze, układ rozprawy oraz wkład Autorki, a także opisuje otwarte dane wykorzystane w analizach.
- **Rozdział Drugi (Podejścia i wyzwania w analizie multiomicznej):** Stanowi przegląd teoretyczny, omawiając obecny stan wiedzy na temat metod statystycznych i uczenia maszynowego w analizie danych multiomicznych. Wyszczególnia etapy analizy oraz identyfikuje kluczowe wyzwania związane z charakterystyką tych danych.
- **Rozdział Trzeci (Faza badawczo-rozwojowa):** Prezentuje eksperymenty przeprowadzone w ramach projektów EPISTOP i EPIMARKER, które stanowiły fazę badawczo-rozwojową pracy. Szczegółowo omawia cele analiz, przygotowanie i eksplorację danych, analizę statystyczną oraz podejścia do modelowania predykcyjnego wraz z uzyskanymi wynikami.
- **Rozdział Czwarty (Metody selekcji cech w analizie multiomicznej):** Skupia się na wynikach eksperymentów dotyczących wyboru optymalnych metod selekcji cech w kontekście danych multiomicznych. Przedstawia przegląd literatury, opis przeprowadzonych eksperymentów, ich wyniki oraz dyskusję wraz z rekomendacjami.
- **Rozdział Piąty (Metody uczenia zespołowego adresujące braki danych):** Omawiane są tu wyniki badań nad zastosowaniem technik uczenia zespołowego w celu rozwiązania problemu niekompletności danych omicznych. Rozdział zawiera przegląd literatury, opis metodologii, szczegóły eksperymentów oraz dyskusję wyników i wnioski.
- **Rozdział Szósty (Konstrukcja potoku przetwarzania - playOmics):** Opisuje autorski potok przetwarzania danych o nazwie playOmics, który został opracowany w odpowiedzi na wyzwania zidentyfikowane we wcześniejszych fazach badań. Zawiera szczegółowy opis etapów przetwarzania, dostępnych funkcji, przykładowe zastosowanie oraz porównanie wyników działania potoku z innymi algorytmami na danych otwartych.

- **Rozdział Siódmy (Implementacja playOmics w projekcie DIPGen):** Poświęcony jest praktycznemu wdrożeniu opracowanego potoku playOmics w ramach badania klinicznego DIPGen, realizowanego w systemie SAMGEN. Przedstawia cel i zakres analizy w tym projekcie oraz sposób przygotowania danych.
- **Rozdział Ósmy (Podsumowanie):** Stanowi końcowe podsumowanie całej rozprawy doktorskiej, prezentując syntezę najważniejszych wniosków płynących z przeprowadzonych badań. Obejmuje również ocenę wdrożeniowego znaczenia pracy oraz jej wkładu w rozwój dziedziny bioinformatyki.

Do pracy został również dołączony imponujący wykaz dorobku naukowego Doktorantki. Rozprawa zawiera 30 szczegółowych rycin/rysunków (w tym dwa w suplemencie) oraz 12 tabel, które co do zasady są jasne i czytelne. Piśmiennictwo obejmuje 138 co do zasady dobrze dobranych i aktualnych pozycji.

Autorka bardzo dokładnie omawia obecny stan wiedzy co świadczy o dużym odczycie i dobrym przygotowaniu do podjęcia zagadnień poruszanych w dalszych częściach pracy. Na wyróżnienie zasługuje fakt, że pomimo ogromu zaprezentowanego materiału czytelnik jest przeprowadzany przez kolejne rozdziały w bardzo umiejętny sposób. Cały układ pracy jest bardzo dobrze przemyślany przez co czytelnik w żadnym momencie nie traci z oczu celu pracy. Ogrom prezentowanych wyników nie wywołuje uczucia przytłoczenia czy zagubienia, a wręcz dzięki umiejętnemu podaniu wzbudza podziw.

Tezy rozprawy są sformułowane jasno i przystępnie oraz są w pełni poparte danymi zawartymi w poszczególnych rozdziałach rozprawy. Podsumowanie rozprawy jest syntetycznym wykazaniem, że założone w pracy cele zostały osiągnięte.

Jeśli chodzi o uchybienia, to co do zasady nie podoba mi się w jaki sposób analiza była przeprowadzona w projektach EPISTOP i EPIMARKER, ale z treści rozprawy wynika, że w wielu miejscach decyzję podejmował „zespół projektowy” a nie Autorka. W wielu wypadkach brakuje uzasadnienia wyborów, w tym tych odnośnie takich a nie innych progów odcięcia itd. Z samej rozprawy wnioskuję, że Autorka też nie była do końca zadowolona z przebiegu projektów i wynikiem tego było powstanie playOmics.

W mojej ocenie znacznym niedopatrzeniem jest fakt, że w sytuacji, gdy jednym z głównych obszarów badania było określenie wpływu sposobu selekcji cech, do analizy nie zostały włączone narzędzia Boruta i MDFS. Zwłaszcza, że Boruta jest cytowana w Rozdziale 2. Oba te narzędzia są przedstawicielami podejścia "all-relevant", do którego znajduje się odniesienie we wprowadzeniu w Rozdziale 4. Myślę, że warto rozszerzyć playOmics o te narzędzia.

Podsumowując, przedstawiona do oceny praca doktorska stanowi bardzo wartościowe uzupełnienie obecnego stanu wiedzy odnośnie analiza danych multiomicznych w oparciu o techniki uczenia maszynowego i metody statystyczne. Rozprawa uwiadamia z jakimi wyzwaniami można się spotkać w rzeczywistych projektach badawczo-rozwojowych i wdrożeniowych. Jednak Autorka nie poprzestaje na ich uwidocznieniu ale proponuje narzędzie (playOmics), umożliwia przeprowadzenie analizy dyskryminacyjnej (klasyfikacji

binarnej) dla zestawu danych omicznych oraz implementuje kroki przetwarzania niezbędne do przygotowania danych do klasyfikacji, prezentacji i interpretacji wyników. Praca ta w pełni odpowiada warunkom stawianym rozprawom doktorskim oraz wypełnia istotną lukę w obecnym stanie wiedzy. Należy też podkreślić ogrom pracy wykonanej przez Doktorantkę jak również imponujący dorobek naukowy.

Na podstawie powyższej oceny stwierdzam, że wymieniona rozprawa doktorska w pełni odpowiada warunkom stawianym w ustawie Prawo o szkolnictwie wyższym i nauce / Dz. U. z 2022 r. poz. 574, w zakresie nadawania stopni naukowych i na tej podstawie wnoszę do Wysockiej Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr inż. Jagody Głowacka-Walas do dalszych etapów przewodu doktorskiego. Równocześnie chciałbym zgłosić pod dyskusję przyznanie wyróżnienia. Motywacją jest zarówno ogrom jak i wysoka jakość wykonanej pracy ale także bardzo wysoka jakość przygotowanej rozprawy.

Nie mam wątpliwości, że doświadczenie zgromadzone przez Autorkę stawia cały zespół badawczy (i/lub zespół w firmie gdzie Autorka jest zatrudniona) w doskonałej pozycji wśród międzynarodowych grup zajmujących się tą tematyką.

Dr hab. inż. Paweł Piotr Łabaj