

**RADA NAUKOWA DYSCYPLINY  
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ**

zaprasza na  
OBRONĘ ROZPRAWY DOKTORSKIEJ

**mgr inż. Jagody Katarzyny Głowackiej-Walas**

która odbędzie się w dniu **25 czerwca 2025 roku**, o godzinie 12:30 w trybie hybrydowym

Temat rozprawy:

**„ Informatyczna analiza danych multiomicznych w oparciu o techniki uczenia maszynowego i metody statystyczne”**

Promotor: dr hab. inż. Tomasz Gambin, prof. uczelni – Politechnika Warszawska

Recenzenci: dr hab. inż. Aleksandra Gruca, prof. uczelni – Politechnika Śląska

dr hab. inż. Paweł Łabaj – Małopolskie Centrum Biotechnologii Uniwersytetu Jagiellońskiego w Krakowie

dr hab. Bartosz Wilczyński, prof. uczelni – Uniwersytet Warszawski

Obrona odbędzie się w sali nr 116 Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej oraz jednocześnie zdalnie na platformie MS Teams (tryb hybrydowy). Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: dr hab. inż. Piotr Gawrysiak, prof. uczelni email: [piotr.gawrysiak@pw.edu.pl](mailto:piotr.gawrysiak@pw.edu.pl), do dnia 23 czerwca 2025r. Godz. 17:00.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: <https://www.bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Rada-Naukowa-Dyscypliny-Informatyka-Techniczna-i-Telekomunikacja/mgr-inz.-Jagoda-Katarzyna-Glowacka-Walas>

Przewodniczący Rady Naukowej Dyscypliny  
Informatyka Techniczna i Telekomunikacja  
Politechniki Warszawskiej  
**prof. dr hab. inż. Jarosław Arabas**

## STRESZCZENIE

Rozprawa koncentruje się na rozwoju narzędzi informatycznych służących do identyfikacji biomarkerów chorobowych oraz wspomagania diagnostyki chorób poprzez analizę danych multiomicznych z wykorzystaniem metod statystycznych i technik uczenia maszynowego. Motywacją do podjęcia badań w tym obszarze były wyzwania zidentyfikowane w dwóch projektach będących częścią badań klinicznych, EPISTOP i EPIMARKER, których celem było opracowanie modeli predykcyjnych umożliwiających przewidywanie wystąpienia napadów padaczkowych lub lekooporności u dzieci ze stwardnieniem guzowatym oraz identyfikację biomarkerów procesu chorobowego. W wyniku tej analizy wyłoniono zagadnienia wymagające uzupełnienia i generalizacji, co stanowiło podstawę do rozszerzenia zakresu badań z wykorzystaniem danych z otwartych baz danych multiomicznych. Pogłębione badania dotyczyły problemu selekcji cech, gdzie przeanalizowano wpływ różnych metod selekcji na wyniki klasyfikacji oraz opracowano zestaw procedur wspierających użytkownika w wyborze odpowiedniej metody. Następnie przeprowadzono badania nad zastosowaniem algorytmów uczenia zespołowego do rozwiązania problemu braków w danych multiomicznych, analizując dwie metody selekcji modeli pod kątem ich zdolności adaptacji do niekompletnych danych. Zaproponowano modyfikację miary niezgodności dla par klasyfikatorów uwzględniającą obserwacje niesklasyfikowane przez wcześniejsze modele z grupy. Tak zdefiniowana metryka wykazała zdolność do selekcji niewielkiej liczby modeli, które umożliwiają poprawną klasyfikację większości obserwacji, w których występują częściowe braki danych. Wyniki realizacji projektów EPISTOP, EPIMARKER oraz badań nad problemem selekcji cech i uczenia zespołowego zostały wykorzystane w konstrukcji potoku przetwarzania playOmics. W rozprawie opisano szczegóły implementacji, przykładowe zastosowanie oraz porównanie wyników działania potoku do innych metod. Potok playOmics umożliwia uproszczenie zarządzania różnorodnymi danymi omicznymi, przetwarzanie wstępne, budowę i ocenę modeli klasyfikacyjnych oraz identyfikację biomarkerów, jednocześnie dostarczając narzędzia zwiększające interpretowalność analizy i ułatwiające jej odtwarzalność. W ostatniej części rozprawy zaprezentowano implementację opracowanego potoku playOmics w badaniu klinicznym DIPGen, umożliwiającą binarną klasyfikację pacjentów dla zdefiniowanego przez użytkownika celu analizy, na podstawie danych pochodzących z różnych warstw omicznych.

**Słowa kluczowe:** uczenie maszynowe, metody statystyczne, dane multiomiczne, poszukiwanie biomarkerów, diagnostyka chorób, choroby rzadkie

Gliwice, 20.03.2025

## Recenzja rozprawy doktorskiej mgr Jagody Głowackiej-Walas

### *Informatyczna analiza danych multiomicznych*

*w oparciu o techniki uczenia maszynowego metody statystyczne*

Ukończonej na Wydziale Matematyki i Nauk Informacyjnych  
Politechniki Warszawskiej

Pod opieką promotora dr hab. inż. Tomasza Gabina, prof. PW

### **Tematyka i cel pracy, problem badawczy i jego znaczenie**

Przedstawiona mi do recenzji rozprawa doktorska powstała na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej pod kierunkiem Profesora Tomasza Gambina. Praca to powstała w formie tak zwanego doktoratu wdrożeniowego i realizowana była przez Doktorantkę zatrudnioną w firmie Transition Technologies-Science (TTSI).

Tematyka rozprawy obejmuje analizę danych multiomicznych za pomocą metod statystycznych oraz technik uczenia maszynowego. Podstawowy cel analiz związany był z budowaniem modeli predykcyjnych umożliwiających przewidywanie wystąpienia napadów padaczkowych lub lekooporności oraz identyfikację biomarkerów procesu chorobowego u dzieci ze stwardnieniem guzowatym, porównanie metod selekcji cech w analizie multiomicznej oraz zaproponowanie techniki pozwalającej na uzyskanie najlepszego zestawu klasyfikatorów w warunkach niekompletności danych.

Rozwój bioinformatyki jest napędzany przez dynamiczny postęp w technologiach wysokoprzepustowych, które znacząco zwiększają możliwości przetwarzania próbek. Skutkuje to generowaniem ogromnych wolumenów danych, wymagających zaawansowania specjalistycznych algorytmów do ich integracji i analizy. Analiza danych multiomicznych odgrywa kluczową rolę w zrozumieniu złożonych procesów biologicznych, ponieważ umożliwia integrację informacji z różnych warstw molekularnych, takich jak genomika, transkryptomika, proteomika i metabolomika. Dzięki temu podejściu możliwe jest uzyskanie szerszego

Politechnika Śląska  
Wydział Automatyki, Elektroniki i Informatyki  
Katedra Sieci i Systemów Komputerowych  
ul. Akademicka 16, pok. 414, 44-100 Gliwice  
+48 32 237 21 51 / +48 237 27 33  
Aleksandra.gruca@polsl.pl

obrazu funkcjonowania organizmu i poznanie kluczowych ścieżek molekularnych związanych z badanymi procesami co pozwala na identyfikację nowych biomarkerów i przez to przyczynia się do rozwoju medycyny spersonalizowanej i precyzyjnej. W przypadku chorób rzadkich, gdzie tradycyjne metody diagnostyczne często zawodzą, analizy multiomiczne stanowią alternatywne narzędzie do odkrywania nowych mechanizmów chorobotwórczych i poprawy diagnostyki, a tym samym poprawy jakości życia pacjentów.

Mimo trwającego już ponad dwie dekady rozwoju metod maszynowego uczenia oraz pojawienia się coraz większej liczby dostępnych danych pochodzących z różnych „omik”, efektywna integracja tych danych pozostaje wyzwaniem. Techniczne i eksperymentalne błędy systematyczne, wprowadzone już na etapie zbierania danych, związane z różnymi protokołami eksperymentalnymi przeprowadzanymi w różnych laboratoriach, wymagają zastosowania odpowiednich technik normalizacyjnych dla heterogenicznych danych. Liczebność próbek, znacznie przekraczająca liczbę badanych pacjentów, oraz związana z tym „klątwa wielowymiarowości”, nieodłączny w analizie danych multiomicznych problem występowania brakujących danych, a także kwestie związane z interpretowalnością wyników zwracanych przez algorytmy powodują, że analiza danych multiomicznych, a w szczególności wykorzystywanie jej w praktyce klinicznej, wciąż pozostaje rzadkością.

W niniejszej rozprawie pani mgr inż. Jagoda Głowcka-Walas podejmuje problematykę niektórych z wymienionych powyżej wyzwań analizy danych multiomicznych. Przedstawione w rozprawie metody koncentrują się na poszukiwaniu biomarkerów przy analizie danych z pojedynczej omiki, opracowanie metod wspomagających wybór algorytmu selekcji cech, wykorzystanie technik uczenia zespołowego do analizy zbiorów danych z brakującymi wartościami oraz stworzeniu narzędzia PlayOmics umożliwiającego budowę modeli predykcyjnych. W kontekście coraz większych możliwości eksperymentalnych pozwalających na analizę różnych omik oraz występującej luki między stosowaniem wyników zintegrowanej analizy danych w zastosowaniach klinicznych, wybrany problem badawczy jest istotny i aktualny oraz bez wątpienia wpisuje się w obecne trendy w dziedzinie bioinformatyki oraz biologii obliczeniowej.

## Charakterystyka rozprawy

Przedstawiona praca doktorska składa się ośmiu rozdziałów.

Pierwszy rozdział stanowi wprowadzenie do rozprawy i koncertuje się przedstawieniu cel i zakres pracy, tezy pracy, wkład autorki w powstałe rozwiązania oraz dane analizowane w ramach doktoratu.

Rozdział drugi zatytułowany jest „**Podejścia i wyzwania w analizie multiomicznej**”. W rozdziale tym, stanowiącym wprowadzenie do tematyki pracy, doktorantka przedstawia kluczowe aspekty związane z przetwarzaniem i interpretacją danych multiomicznych, ogólny kontekst analiz multiomicznych oraz opisuje rodzaje danych omicznych wraz z metodami ich przetwarzania. Następnie omawia zastosowania analiz multiomicznych w medycynie i biologii. Kolejne podrozdziały skupiają się na etapach przetwarzania danych, w tym na przygotowaniu i eksploracji danych, redukcji wymiarowości oraz modelowaniu, które obejmuje analizę statystyczną, uczenie maszynowe, ocenę i walidację modeli, a także ich zastosowania w praktyce klinicznej i interpretowalność. Przedstawione są główne wyzwania analizy, takie jak duży rozmiar i złożoność danych, efektywna integracja danych omicznych, mała liczba próbek (szczególnie w kontekście

chorób rzadkich), problem brakujących danych oraz niezbalansowanie klas. Rozdział kończy się omówieniem dobrych praktyk i rekomendacji dla analizy danych multiomicznych, które mają na celu poprawę jakości i efektywności badań.

Rozdział trzeci „**Faza badawczo-rozwojowa**” opisuje wyniki badań wykonanych w ramach projektu EPISTOP oraz EPIMARKER. W projekcie EPIMARKER doktorantka pełniła rolę kierownika projektu ze strony firmy TTSL. W ramach projektu EPISTOP analizowano dane transkryptomiczne, proteomiczne, wybrane dane metabolomiczne, miRNA oraz analizowano DNA. W ramach prac porównano klasyfikatory XGBoost, Lasso oraz regresję logistyczną. Finalnie analizy porównawcze wykonano dla klasyfikatorów w oparciu o regresję logistyczną dla wszystkich możliwych jedno- dwu- i trzelementowych kombinacji zmiennych. Z kolei opisane w doktoracie badania wykonane w ramach projektu EPIMARKER, gdzie analizowano dane kliniczne, proteomiczne, miRNA oraz RNA. Zastosowano tu dwa podejścia: budowa jednego dużego modelu wykorzystującego wszystkie cechy, które pozostały po etapie filtrowania. To podejście przyniosło niezadowolającą jakość predykcji. Budowa wielu małych modeli regresji logistycznej (z dwoma lub trzema zmiennymi wejściowymi). To podejście zostało przyjęte w oparciu o obiecujące wyniki uzyskane przy użyciu małych modeli w projekcie EPISTOP. W obydwu projektach, dużym wyzwaniem była niewielka liczba pacjentów (34 i 65) w obydwu projektach.

W kolejnym rozdziale pt. „**Metody selekcji cech w analizie multiomicznej**” doktorantka koncentruje się na problemie, jakim jest wysoka wymiarowość danych multiomicznych, gdzie liczba cech często znacznie przekracza liczbę próbek ( $p \gg n$ ). W rozdziale tym przedstawiono wyniki dla analiz porównania metod selekcji cech dla pięciu otwartych zbiorów danych wykonanych w ramach pięciu eksperymentów: (1) porównanie wybranych metod selekcji cech (jednowymiarowe: AUC, IG; wielowymiarowe: MRMR; wbudowane: RF, Lasso; oraz DIABLO) w różnych ustawieniach. (2) Analiza pokrywania się oraz typów cech omicznych wybranych przez różne metody, aby zrozumieć ich stabilność i tendencje do preferowania określonych typów danych. (3) Badanie czy uwzględnienie niekompletnych obserwacji w zbiorach danych może poprawić wyniki klasyfikacji, ze szczególnym uwzględnieniem metod jednowymiarowych (AUC, IG), które radzą sobie z brakującymi danymi. (4) Analiza wpływu redukcji liczby próbek na wydajność różnych metod selekcji cech. (5) Porównanie czasu obliczeniowego wymaganego przez różne metody selekcji cech. W podsumowaniu, stwierdzono, że nie istnieje jedna uniwersalna metoda selekcji cech i zawsze należy wybierać metodę selekcji cech w oparciu o charakterystykę zbioru. Równocześnie zostało stworzone narzędzie symulacyjne narzędzia wspomagające proces wyboru właściwej metody selekcji cech.

Rozdział 5 „**Metody uczenia zespołowego adresujące braki danych**” poświęcony jest metodom uczenia zespołowego, które są stosowane w celu rozwiązania problemu braków danych w analizie multiomicznej. W rozdziale analizowano dwa główne podejścia do wyboru zespołu klasyfikatorów: statyczne i dynamiczne. Problem selekcji statycznej polega na wyborze stałego zestawu klasyfikatorów przed fazą predykcji. Doktorantka zaproponowała tu modyfikację powszechnie stosowanej metryki mierzącej niezgodność między klasyfikatorami. Kluczowym pomysłem było traktowanie brakujących predykcji jako formy niezgodności między modelami. W ten sposób proces selekcji ma na celu znalezienie kolejnych klasyfikatorów do zespołu, które mogą dokonywać predykcji dla obserwacji, których poprzednie modele nie były w stanie przetworzyć z powodu brakujących danych. Selekcja dynamiczna polega na wyborze podzbioru klasyfikatorów z puli dla każdej nowej obserwacji na podstawie jej charakterystyki, w tym wzorca brakujących danych.

W szczególności do oceny próbek wykorzystywane są tylko takie modele, które dla danej wykorzystują te dane, które dostępne są dla danej próbki. Eksperymenty w rozdziale piątym przeprowadzono głównie na rozszerzonej wersji zbioru danych BRCA z TCGA, który zawierał dane dotyczące zmienności liczby kopii ( $BRCA_R$ ). Aby symulować scenariusze z większą liczbą brakujących danych, zarówno zbiory walidacyjne, jak i testowe poddano losowemu maskowaniu, zapewniając, że co najmniej 20% wierszy dla każdej omiki zawiera brakujące wartości. Ocena zaproponowanych metod koncentrowała się na liczbie brakujących predykcji w zbiorze testowym oraz dokładności dokonanych predykcji. Wyniki wykazały, że zaproponowane podejście do selekcji statystycznej, szczególnie z wykorzystaniem poprawionej metryki  $Miss_{corr}$ , było skuteczne w minimalizowaniu brakujących predykcji przy zachowaniu dobrej dokładności klasyfikacji przy mniejszym rozmiarze zespołu klasyfikatorów w porównaniu z poleganiem wyłącznie na jakości predykcji

Przedostatni rozdział rozprawy pt. „**Konstrukcja potoku przetwarzania – playOmics**” zawiera opis pakietu R o nazwie playOmics. Narzędzie to zostało stworzone, aby rozwiązać wyzwania napotykane w analizie danych multiomicznych, wykorzystując doświadczenia z projektów EPISTOP i EPIMARKER. Głównym celem playOmics jest uproszczenie procesy analizy złożonych danych omicznych, umożliwienie budowy skutecznych modeli predykcyjnych, jednocześnie priorytetyzując interpretowalność, łatwość użycia i powtarzalność analiz w celu identyfikacji markerów prognostycznych. Etapy analizy multiomicznej, które umożliwia narzędzie to przygotowanie danych, selekcja cech, trenowanie i walidacja modeli. W odniesieniu do wyników przedstawionych w ramach pracy doktorskiej w narzędziu zaimplantowano metodę tworzenia niewielkich  $m$ -czynnиковych modeli regresji logistycznej, narzędzie symulacyjne wspomagające selekcję cech, oraz narzędzie wyboru modeli do analizy danych brakujących przy użyciu metryki  $Miss_{corr}$ . Działanie pakietu playOmics zaprezentowano dla zbiorów danych BRCA oraz THCA.

Recenzowana praca stanowi doktorat wdrożeniowy, dlatego istotnym elementem oceny jest tu czy przedstawione rozwiązania zostały zaimplantowane w postaci narzędzia i wykorzystane w ramach istniejących, praktycznych projektów. W rozdziale 6, omówionym powyżej, opracowane metody zostały zaimplementowane w pakiecie PlayOmics, w rozdziale 7 „**Implementacja playOmics w projekcie DIPGen**” opisano wykorzystanie narzędzia playOmics jako jednego z modułów systemu analizy danych wykorzystywanego w projekcie DIPG koordynowanym przez Centrum Zdrowia Dziecka w Warszawie.

Ostatni rozdział pracy zwięźle opisuje wnioski z wykonanych analiz oraz kierunki dalszych prac.

## Opinia o rozprawie

Należy podkreślić, iż problem, którego podjęła się rozwiązać Doktorantka jest niewątpliwie ważny i wpisujący się w trendy najnowszych badań w dziedzinie. Analiza danych multiomicznych, mimo dużego potencjału i zaawansowanych technologii pomiarowych, wciąż nie jest szeroko stosowana w praktyce klinicznej ze względu na złożoność i heterogeniczność danych, trudności w integracji różnych warstw omicznych oraz problemy z brakującymi danymi. Dodatkowo, analiza ta wymaga dużych zasobów obliczeniowych, a modele predykcyjne często są trudne do interpretacji klinicznej, co ogranicza ich praktyczne zastosowanie. Wreszcie, dla wielu chorób brakuje wystarczającej liczby próbek do wiarygodnej walidacji i generalizacji wyników.

Przedstawiona mi do oceny praca doktorska adresuje niektóre z wyzwań związanych z analizą danych multiomicznych. W rozdziale trzecim autorka mierzy się z praktycznymi problemami analizy danych multiomicznych w chorobach rzadkich, związanymi z częstą w realnych projektach koniecznością analizy niewielkiej liczby próbek pacjentów. Rozdział czwarty opisuje opracowanie narzędzia symulacyjnego, które wspiera użytkowników w wyborze odpowiednich metod selekcji cech dla ich specyficznych zbiorów danych multiomicznych. W rozprawie przedstawiono wpływ różnych metod selekcji cech na wyniki klasyfikacji oraz wskazówki dotyczące ich wyboru w zależności od charakterystyki danych. Rozwiązanie to odpowiada na wyzwania związane z wysoką wymiarowością danych oraz brakiem jednej, uniwersalnie optymalnej metody w analizie multiomicznej. Rozdział piąty przedstawia badania dotyczące zastosowania algorytmów uczenia zespołowego w celu rozwiązania problemu brakujących danych w zbiorach danych multiomicznych. Zaproponowano zmodyfikowaną metrykę różnorodności klasyfikatorów, uwzględniającą szczególnie obserwacje niezaklasyfikowane z powodu brakujących danych. Opisane w tym rozdziale prace stanowią wkład w trwające wyzwanie związane z analizą niekompletnych zbiorów danych multiomicznych. Stworzone w ramach pracy narzędzie playOmics zostało zaprojektowane z silnym naciskiem na zwiększenie interpretowalności analiz oraz zapewnienie reprodukowalności. Zaimplementowanie graficznego interfejsu użytkownika (GUI), szczegółowe rejestrowanie przebiegu eksperymentów oraz generowanie włączenie do pakietu narzędzi pozwalających na interpretowanie wyników modeli (wartości SHAP) bezpośrednio odpowiadają na potrzebę bardziej przejrzystych i zrozumiałych analiz multiomicznych, gdzie reprodukowalność jest kluczowa dla walidacji i dalszego wykorzystania wyników badań. Przy ocenie pracy należy też zwrócić uwagę na praktyczne aspekty przedstawionych badań.

W ostatniej części rozprawa prezentuje zastosowanie narzędzia playOmics (jako elementu aplikacji GeneSense) w projekcie DIPGen do klasyfikacji binarnej pacjentów na podstawie danych multiomicznych. Demonstruje to rzeczywistą przydatność i zastosowanie playOmics w środowisku badań klinicznych, szczególnie w kontekście choroby rzadkiej. Ponadto, metodologie klasyfikacji oparte o budowę niewielkich *m*-czynnиковych modeli regresji logistycznej zaimplementowane w playOmics były wcześniej wykorzystywane w projektach EPISTOP i EPIMARKER, co jest kolejnym argumentem wskazującym, że zaprojektowane przez panią mgr. inż. Jagodę Głowacką-Walas metody wykorzystywane są do poszukiwania nowych biomarkerów w projektach naukowych.

Spis literatury składa się ze 138 pozycji. Jak na pracę doktorską, nie jest to duży przegląd, jednak oddaje on aktualny stan wiedzy w zakresie, którego dotyczy rozprawa. Wiele z cytowanych pozycji to prace najnowsze co stanowi dowód na to, iż Doktorantka orientuje się w aktualnych badaniach prowadzonych w tematyce rozprawy. Świadczy to również o tym, że problem, którego podjęła się rozwiązać oraz wykorzystane w tym celu metody wpisują się w najnowsze trendy badań związanych z analizą i integracją danych multiomicznych.

Pozytywnie też oceniam fakt, iż kody źródłowy metod narzędzia playOmics został udostępniony środowisku naukowemu na platformie GitHub wraz z konfiguracją pozwalającą na uruchomienie narzędzia w kontenerze.

Podsumowując, uważam, iż w swojej pracy Doktorantka zastosowała odpowiednie metody badawcze w celu rozwiązania problemu oceny efektywności różnych metod statystycznych oraz technik uczenia maszynowego w analizie danych multiomicznych w obszarach selekcji cech oraz brakujących danych. Cel rozprawy został osiągnięty. Zastosowane w pracy podejścia do przetwarzania danych wskazują na dobrą znajomość przez Doktorantki nowoczesnych i efektywnych metod stosowanych w dziedzinie biologii obliczeniowej.

## Uwagi krytyczne i dyskusyjne

Należy podkreślić, że problem, który podjęła się rozwiązać mgr inż. Jagoda Głowacka-Walas jest niewątpliwie ważny i trudny i na początku tej części chciałbym podkreślić, że nie znalazłam w przedstawionych wynikach żadnych zasadniczych błędów merytorycznych. Wszystkie poniższe uwagi wynikają z chęci podjęcia dyskusji i dialogu na temat niektórych aspektów pracy. Uwagi te nie obniżają mojej pozytywnej oceny pracy.

Za istotny mankament pracy uważam w części teoretycznej raczej dość słaby przegląd literatury zarówno dla metod selekcji cech jak i metod radzenia sobie z brakującymi danymi w analizie danych multiomicznych, a także brak szerszego porównania zaproponowanych rozwiązań z metodami *state-of-the-art*. Przynajmniej w przypadku prac związanych z omówieniem metod selekcji cech Doktorantka jako jedną z metod do porównania wybrała metodę DIABLO, która dedykowana jest do analizy danych multiomicznych. Natomiast w rozdziale związanym z podejściem do problemu brakujących wartości, zaproponowana metoda nie jest porównywana do żadnych innych znanych z literatury metod radzenia sobie z brakującymi danymi, a w pracy nie podjęto nawet dyskusji teoretycznej porównującej zaproponowane podejście z istniejącymi metodami. Istnieje wiele prac przeglądowych, np. praca przeglądowa Flores et al. [10.3389/frai.2023.1098308], czy praca Hornung et. al. [doi.org/10.1002/wics.1626], obie z nich cytowane w rozprawie, które omawiają szereg metod wykorzystywanych do radzenia sobie z problemem brakujących danych.

W przypadku wielokrotnego budowania modeli istnieje ryzyko, że jeden z kilku tysięcy testowanych modeli przypadkowo dopasuje się do dodanych dając nadmiernie optymistycznie wyniki. W pracach opisanych w rozdziale trzecim zastosowano odpowiednie metody mitygacji takiej sytuacji przez zastosowanie symulacji Monte Carlo oraz wykorzystywanie oddzielnego zbioru walidacyjnego. Czy stworzone w ramach doktoratu narzędzie PlayOmics zawiera mechanizmy wymuszające na użytkowniku stosowanie metod przeciwdziałania przeuczeniu modeli?

W pracy podkreślono ważność interpretacji wyników zwracanych przez modele. W pracy możliwość interpretowalności została zrealizowana z jednej strony poprzez zaproponowanie interpretowalnych modeli regresji, a z drugiej przez zaimplementowanie metody SHAP w narzędziu PlayOmics. Czy w analizach przeprowadzonych i przedstawionych w ramach pracy doktorskiej dla zbiorów BRCA i THCA uzyskane wyniki pokrywały się? Jeśli nie, co mogło być przyczyną różnic?

### Uwagi redakcyjne:

W podrozdziale 4.6 warto by było umieścić rekomendacje w zbiorczej tabeli, ułatwiłoby to podsumowanie i syntezę wniosków uzyskanych w ramach badań przeprowadzanych w rozdziale 4.

Dobrze byłoby zadbać o bardziej staranne przygotowanie rysunków do pracy. Część z pokazanych na nich wyników jest trudno zinterpretować (np. rys 4.5), dla niektórych „ucięto” opisy osi (np. rys 6.4 D, rys 6.7), a część informacji jest nieczytelna (rys 6.4 C i F)

## Podsumowanie

Pani mgr inż. Jagoda Głowacka-Walas przedstawiła rozprawę doktorską rozwiązującą aktualny problem naukowy, która przyczyni się do rozwoju reprezentowanej dyscypliny naukowej. Rozprawa zawiera oryginalne rozwiązanie problemu naukowego, a Doktorantka wykazała, że zarówno posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja jak i umiejętność prowadzenia pracy naukowej.

Pani mgr inż. Jagoda Głowacka-Walas jest współautorką 9 publikacji naukowych, w tym w trzech z nich występuje 3 jako członek konsorcjum EPISTOP. Jedna z tych publikacji została opublikowana w prestiżowym czasopiśmie Nature Communications, gdzie pani mgr inż. Jagoda Głowacka-Walas posiada pierwsze autorstwo współdzielone. Jest to niewątpliwie duże osiągnięcie. Wg bazy danych Scopus większość z czasopism, w których publikuje doktorantka to czasopisma z dziedziny medycyny, neurologii, pediatrii co wskazuje na aplikacyjny charakter przeprowadzonych prac i ich silne powiązanie z projektem. Ponadto wyniki prac zostały zaimplementowane w postaci narzędzia i wykorzystane do analizy danych w projekcie naukowym. W doktoratach wdrożeniowych, biorąc pod uwagę charakter prac, które koncertują się bardziej na aspekcie wdrożenia produktu, zwyczajowe wymagania publikacyjne są zwykle niższe niż w przypadku doktoratów tradycyjnych. Dlatego też bardzo wysoko oceniam fakt, iż Doktorantka podjęła wysiłek publikacyjny zakończony sukcesem w tak dobrych czasopismach.

Biorąc pod uwagę ocenę przedstawioną w niniejszej recenzji, stwierdzam, że przedstawiona do oceny praca doktorska w pełni odpowiada warunkom określonym w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2024 r. poz. 1571 z późn. zm.) i na tej podstawie wnoszę do Wysockiej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr inż. Jagody Głowackiej -Walas do dalszych etapów przewodu doktorskiego.

dr inż. hab. Aleksandra Gruca, prof. PŚ.

## **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 odczuciu 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), umożliwiające 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



# UNIwersytet Warszawski

Instytut Informatyki  
Uniwersytet Warszawski  
ul. Banacha 2  
02-097 Warszawa  
POLSKA

dr hab. Bartosz Wilczyński  
profesor uczelni  
Phone: +(48 22) 5544 577  
Fax: +(48 22) 5544 400  
e-mail: bartek@mimuw.edu.pl

Warszawa, 4. maja 2025 r.

Recenzja rozprawy doktorskiej pt. „Informatyczna analiza danych multiomicznych w oparciu o techniki uczenia maszynowego i metody statystyczne” przedstawionej przez mgr inż. Jagodę Głowacką – Walas

Recenzja niniejsza została sporządzona na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej zgodnie z wymogami ustawy dotyczącej procedur nadawania stopnia doktora. Recenzja ta składa się najpierw ze skrótego opisu merytorycznej zawartości rozprawy, następnie zawiera moje uwagi krytyczne dotyczące cech rozprawy, które stanowią pewne usterki czy niedociągnięcia i zakończona jest podsumowaniem.

## Opis rozprawy

Praca złożona przez p. Głowacką-Walas jest obszerna, zawiera wyniki opublikowane już wcześniej w kilku publikacjach. Pomijając rozdział 1, stanowiący wprowadzenie do rozprawy, rozprawa składa się z kilku głównych części (jak opisuje to sama autorka):

- Wstęp metodologiczny w rozdziale 2.
- Wyniki prac w projektach EPISTOP i EPIMARKER w rozdziale 3.
- Wyniki eksperymentów dotyczących metod selekcji cech i uczenia zespołowego w rozdziałach 4-5
- Model narzędzia PlayOmics w rozdziale 6 i przykładowa implementacja potoku w rozdziale 7.

Z punktu widzenia wkładu autorki do badań naukowych, niewątpliwie jej udział w konsorcjum EPISTOP i projektach związanych z tym dużym przedsięwzięciem naukowym zaowocowały najbardziej widocznymi publikacjami. Z drugiej strony, niestety te wyniki budzą największe wątpliwości dotyczące metodologii. Z kolei wyniki przedstawione w późniejszych rozdziałach, czy to wyniki symulacji i eksperymentów z rozdziałów 4-5, czy implementacja narzędzia playOmics w rozdziałach 6-7, cechują się dużo lepszym warsztatem obliczeniowym, choć niestety nie mają takiego wpływu na środowisko naukowe.

Z mojego punktu widzenia, wszystko co w rozprawie jest napisane, wskazuje, że autorka miała styczność z dużym projektem i rzeczywistymi danymi klinicznymi, co niestety - mimo niewątpliwie dużego wysiłku autorki włożonego w konstrukcję efektywnych klasyfikatorów, nie doprowadziło do budowy modeli przełomowych dla predykcji lekooporności czy stopnia dolegliwości ataków w badanych przypadkach. Niemniej, doświadczenie zdobyte przez autorkę w tych projektach zaowocowało wykonaniem przez nią kolejnych eksperymentów (opisanych w rozdziałach 4-5) i przygotowanie narzędzia Playomics, które może posłużyć kolejnym badaczom. Trudno jednoznacznie ocenić, czy możliwe byłoby uzyskanie lepszych wyników klasyfikacji w projekcie EPISTOP, jednak biorąc pod uwagę skąpość dostępnych danych i trudność niektórych z problemów, wydaje się to bardzo prawdopodobne, że dla tych danych nie byłoby to możliwe.

## Uwagi krytyczne

### Istotna wątpliwość dotycząca wyboru modeli predykcyjnych na podstawie testów permutacyjnych

Autorka przedstawia rozwiązanie korzystające z “małych” modeli regresji logistycznej w projekcie EPISTOP. Wykorzystuje tam “małe” modele regresji logistycznej i ocenia je na podstawie testów permutacyjnych. W szczególności na rysunku 3.6, jeśli dobrze rozumiem, przedstawia rozkład miary MCC na zbiorach testowych dla bardzo wielu modeli i zwraca uwagę na dominującą wartość 0, oznaczającą niską jakość predykcji. Później wykonuje testy na permutowanych danych i w ten sposób wybiera próg, aby co najwyżej 5 procent klasyfikatorów losowych było lepszych niż te wybrane ostatecznie. Nie jestem przekonany, czy to jest wystarczająca metoda. W szczególności, gdyby okazało się, że rozkład testowych wartości MCC dla permutowanych danych jest podobny do tego na wykresie 3.6, należałoby uznać, że zmienne objaśniające nie wnoszą istotnej informacji. zmiennej objaśnianej i nie można byłoby wybierać po prostu 5 procent najlepszych losowych klasyfikatorów. Fakt, że autorka spośród 22100 modeli, które mają być lepsze niż 0.05 losowych wybiera 100 modeli, sugeruje, że w istocie tych modeli jest około 0.05 co sugerowałoby, że mamy do czynienia z niemalże losowymi klasyfikatorami.

Chciałbym, aby autorka jakoś odniosła się do tej wątpliwości. Tzn. uważam, że nawet jeżeli w ostatecznym rozrachunku wynik tych badań jest negatywny (tzn. żaden z tych modeli nie jest lepszy niż to czego spodziewamy się w modelu zupełnie losowym, bez

związku pomiędzy zmiennymi objaśniającymi a objaśnianymi), to nie przekreśla całości dorobku doktorantki, ale jednak chciałbym wiedzieć, czy tam naprawdę jest jakaś nielosowa zależność, czy jednak ostatecznie zwrócone modele w zasadzie nie różnią się od tego, czego moglibyśmy się spodziewać, gdybyśmy takie dane wygenerowali zupełnie losowo, bez związku pomiędzy zmienną objaśnianą i predyktorami.

## Drobniejsze uwagi natury redakcyjnej i stylistycznej

W rozdziale 2, autorka opisując działy takie jak genomika, transkryptomika i proteomika, czyni to w sposób niezmiernie skrótowy. Uwzględnia tu też wyłącznie typy danych, które później sama analizuje (pomijając np. dane takie jak ChIP-Seq czy Hi-C). Ta część mogłaby zyskać na bardziej przekrojowym potraktowaniu materiału.

W rozdziale 3, rysunek 3.3 przedstawia krzywe ROC dla 3 modeli: z 1 zmienną, wszystkimi zmiennymi i “optimalny” z 6 zmiennymi. Niestety, krzywa ROC dla “optimalnego” modelu jest wyraźnie gorsza niż dla modelu z 1 zmienną. Autorka nie komentuje tego faktu, choć powinna go zauważyć. Rozumiem, że model wybrany poprzez optymalizację MCC może się nie sprawdzać, ale jednak wymagałoby to jakiegoś komentarza.

Autorka często używa sformułowania “adresować” w znaczeniu będącym kalką angielskiego słowa “address”, czyli “znaleźć rozwiązanie” czy “odpowiedzieć na wątpliwość”, co nie jest zgodne z polskim znaczeniem tego słowa, ale pojawia się coraz częściej w potocznej polszczyźnie.

## Podsumowanie

Podsumowując, praca mgr inż. Jagody Głowackiej-Walas podejmuje istotne problemy w dziedzinie bioinformatyki. Mimo pewnych niedociągnięć i usterek, o których pisałem wcześniej, praca stanowi niewątpliwie wkład do dyscypliny informatyka w dziedzinie nauk technicznych i świadczy o tym, że autorka uzyskała poziom wiedzy i samodzielności naukowej oczekiwany na etapie doktoratu. W związku z tym uważam, że **rozprawa doktorska spełnia ustawowe wymagania wobec prac doktorskich** i może zostać skierowana do kolejnych etapów przewodu doktorskiego.

Z poważaniem,



Bartosz Wilczyński