Dr hab. inż. Grzegorz Dudek, prof. PCz
Katedra Automatyki, Elektrotechniki i Optoelektroniki
Wydział Elektryczny
Politechnika Częstochowska
Al. Armii Krajowej 17
42-200 Częstochowa

Częstochowa, dn. 3 stycznia 2023 r.

R E C E N Z J A

rozprawy doktorskiej mgr inż. Katarzyny Frątczak
pt. *Dimensionality reduction and feature selection methods in the problems of grouping and classification of molecular imaging data*

Formalną podstawą opracowania recenzji jest pismo Przewodniczącego Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej, prof. dr hab. inż. Marka Gzika, z dnia 03.11.2022 r. Oceny rozprawy doktorskiej dokonano według kryteriów określonych w ustawie z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Promotorami rozprawy doktorskiej są prof. dr hab. inż. Joanna Polańska oraz dr hab. Monika Pietrowska, prof. NIO.

Charakterystyka rozprawy

Rozprawa napisana jest w języku angielskim, liczy 118 stron. Składa się z dziewięciu rozdziałów, bibliografii zawierającej 40 pozycji, spisów tabel i rysunków, listy skrótów oraz informacji na temat finansowania. Praca poprzedzona jest streszczeniami w języku angielskim i polskim.

Tezy pracy są następujące:

1. *Adaptive baseline correction and Gaussian mixture model feature extraction significantly reduce the data dimensionality without losing biological information.*
2. *Introducing the intelligent construction of training datasets utilising information about the heterogeneity of tissue subtypes improves the classification quality.*
3. *The proposed method of feature selection using adaptive scoring allows for covering the heterogeneity of the cancer dataset and finding the robust final signature of cancer.*

We wstępie rozprawy Autorka przedstawiła motywację, zawarła cel i tezy pracy oraz opisała zawartość poszczególnych rozdziałów.

W rozdziale drugim, *Biological background*, omówiono zagadnienia biologii molekularnej, histopatologii, proteomiki klinicznej i lipidomiki.

Rozdział trzeci, *Mass spectrometry,* stanowi wprowadzenie do spektrometrii masowej. Opisano techniki spektrometrii, a w szczególności technikę wykorzystywaną w badaniach, *matrix-assisted laser desorption and ionization*, MALDI-TOF. Opisano zastosowanie tej techniki do identyfikacji peptydów jako biomarkerów nowotworowych oraz obrazowania przestrzennego materiału biologicznego.

W rozdziale czwartym, *Materials*, opisano dane biologiczne, które są przedmiotem badań. Pierwszy zbiór danych wykorzystywany w badaniach obrazowania peptydów i lipidów zawiera cztery zestawy

próbek nowotworów głowy i szyi. Drugi zbiór wykorzystywany w badaniach obrazowania na podstawie mikromacierzy tkankowych zawiera sześć zestawów próbek różnych nowotworów.

W rozdziale piątym, *Pre-processing of molecular imaging data*, opisano procedurę przygotowania danych do klasyfikacji. Obejmuje ona „czyszczenie" danych – detekcję i usuwanie błędów, szumów i obserwacji odstających, „wyrównywanie" widm oraz modelowanie widm za pomocą mieszaniny rozkładów Gaussa.

Rozdział szósty, *Local cancer-dependent signature*, przedstawia etapy budowy systemu klasyfikacyjnego i procedury identyfikacji lokalnych wzorców molekularnych dla różnych typów nowotworów. Opisano metody redukcji wymiarowości danych oraz klasyfikator oparty na regresji logistycznej. Opisano zagadnienia bioróżnorodności pacjentów cierpiących na ten sam typ nowotworu oraz bioróżnorodności wewnątrztkankowej. Zaproponowano dwa podejścia do klasyfikacji dla tych dwóch typów bioróżnorodności. Opisano metody selekcji cech oraz metody radzenia sobie z danymi niezbalansowanymi, a także mierniki jakości klasyfikacji.

Rozdział siódmy, *Global cancer profile*, opisuje porównanie dokładności i sygnatur (wzorców, cech istotnych) klasyfikatorów lokalnych rozpoznających poszczególne typy nowotworów i klasyfikatora globalnego. Porównano sygnatury lokalne z sygnaturą globalną, utworzoną dla wszystkich badanych typów nowotworów.

Rozdział ósmy, *Comparison of peptide and lipid cancer profile*, dotyczy klasyfikacji zbiorów danych wykorzystywanych w badaniach obrazowania peptydów i lipidów. Opisano metodę segmentacji obrazów tkanek na heterogeniczne obszary, selekcję cech istotnych oraz wyniki klasyfikacji.

Badania podsumowano w rozdziale dziewiątym.

Opinia na temat rozprawy

Tematyka rozprawy dotyczy obróbki, grupowania i klasyfikacji danych pozyskanych techniką spektrometrii mas w zastosowaniu do rozpoznawania nowotworów. Obrazowanie molekularne za pomocą spektrometrii mas jest podstawowym narzędziem diagnostyki nowotworów. Złożoność sygnału widma masowego, duża ilość danych dla analizowanych próbek tkankowych oraz wysoka heterogeniczność komórek stanowi wyzwanie zarówno dla klasycznych metod analizy statystycznej jak i metod uczenia maszynowego. Autorka zauważa, że komercyjne narzędzia analizy tkanek na podstawie spektrometrii mas mają poważne ograniczenia w wielopoziomowym przetwarzaniu i analizie wielkoskalowych danych i proponuje algorytmy i narzędzia obróbki, analizy i klasyfikacji wielowymiarowych danych obrazowania molekularnego. Autorka opracowała schemat przetwarzania danych obejmujący adaptacyjną korekcję linii bazowej widma, ekstrakcję cech przy użyciu mieszanin rozkładów Gaussa oraz selekcję cech opartą na adaptacyjnym rankingu. Opracowała metody konstrukcji zbioru uczącego z wykorzystaniem informacji o niejednorodności podtypów tkanek, niwelujące problem niezbalansowania danych. Działania te doprowadziły do identyfikacji sygnatur badanych nowotworów i poprawy jakości ich klasyfikacji. Wyniki badań przyczyniają się do lepszego zrozumienia biologicznych podstaw rozwoju nowotworów i dokładniejszej diagnostyki, co ma ogromne znaczenie w kontekście wzrastającej z roku na rok liczby zachorowań na choroby nowotworowe, które należą do jednej z najczęstszych przyczyn zgonów. Tematykę pracy uznaję za bardzo ważną, społecznie pożyteczną i aktualną w aspekcie jej walorów poznawczych i utylitarnych.

Tytuł rozprawy jest odpowiednio zwarty i komunikatywny. W pełni oddaje najistotniejsze elementy treściowe rozprawy. Motywacja do podjęcia badań przedstawiona przez Autorkę we wstępie pracy jest przekonująca. Cel badań jest jasno postawiony; jest nim opracowanie i implementacja metod przetwarzania danych obrazowania molekularnego – widm masowych komórek, w celu budowy systemu klasyfikującego, rozróżniającego komórki nowotworowe i zdrowe. Tezy pracy są poprawne, oryginalne i jednoznaczne.

W rozdziale drugim Autorka przygotowuje czytelnika do lektury kolejnych rozdziałów przedstawiając kontekst biologiczny. Podkreśla znaczenie obrazowania nowotworów za pomocą spektroskopii masowej, która umożliwia przestrzenną identyfikację profili molekularnych i ich niejednorodności w guzie. Wyjaśnia czym zajmuje się histopatologia i opisuje preparację tkanek do badań histopatologicznych. Charakteryzuje genomikę, transkryptomikę oraz proteomikę. Ta ostatnia gałąź nauk biologicznych, zajmująca się badaniem białek, ich budowy, funkcji i interakcji między nimi, ma kluczowe znaczenie w badaniach podjętych w ramach pracy doktorskiej. Autorka słusznie podkreśla trudności w badaniach proteomu, które związane są z ogromnym zróżnicowaniem cząsteczek białkowych w komórkach. To zróżnicowanie wynika z lokalizacji białka wewnątrz komórki i fazy cyklu komórkowego oraz warunków środowiskowych. W dalszej części Autorka prezentuje zagadnienia lipidomiki istotne z punktu widzenia podjętych badań, czyli identyfikacji i oznaczania ilościowego lipidów na podstawie widm masowych, a następnie budowy modelu klasyfikacyjnego rozróżniającego komórki zdrowe od nowotworowych. Autorka przekonująco uzasadnia te badania tym, że lipidy jako cząsteczki sygnałowe i metaboliczne, mają głęboki wpływ na aktywację immunologiczną i rozwój wielu chorób. Nieprawidłowy metabolizm lipidów jest wspólną cechą komórek nowotworowych, którą można zaobserwować we wczesnych stadiach rozwoju nowotworu, co sprzyja wykorzystaniu badań lipidów do wczesnego wykrywania raka. Ponadto rozwój technologiczny jaki nastąpił w ostatnich latach (wynalezienie techniki „miękkiej jonizacji") umożliwia bezpośrednie wykrywanie lipidów w próbkach tkanek in situ.

Narzędziem wykorzystywanym w badaniach do obrazowania danych molekularnych jest spektrometr mas. Autorka w rozdziale trzecim podaje rys historyczny spektrometrii mas, opisuje zasadę działania i budowę spektrometru. Podaje techniki jonizacji i typy jonów, opisuje analizatory mas, które separują jony w zależności od stosunku ich mas do ładunku oraz detektory, które rejestrują jony przechodzące przez analizator w postaci sygnałów elektrycznych. W dalszej części Autorka opisuje zastosowanie spektrometrii mas w proteomice klinicznej w celu identyfikacji biomarkerów białkowych stosowanych w diagnostyce molekularnej chorób. Zauważa, że spektrometria mas pozwala na nieoperacyjną diagnostykę nowotworów i mikroprzerzutów na etapie „przedklinicznym". W wyodrębnionych podrozdziałach Autorka opisuje szczegółowo: spektrometrię w technologii MALDI-TOF (*Matrix-Assisted Laser Desorption and Ionization - Time Of Flight*), którą używa w swoich badaniach, strategie identyfikacji protein za pomocą spektrometrii mas oraz procedurę obrazowania molekularnego za pomocą MALDI-TOF – od przygotowania próbki do tworzenia obrazów molekularnych 3D pokazujących intensywność jonów w funkcji ich przestrzennego położenia w próbce. Lektura drugiego i trzeciego rozdziału pracy daje poczucie, że Autorka jest dobrze przygotowana do przeprowadzenia badań. Nabyła niezbędną wiedzę teoretyczną z zakresu biologii molekularnej oraz wiedzę techniczną i umiejętności z zakresu spektrometrii mas. Swobodnie porusza się w tych obszarach.

Materiał do badań opisano w rozdziale czwartym. Pierwszy zestaw próbek obejmuje widma masowe dla peptydów i lipidów otrzymane z wycinków tkankowych dla czterech pacjentów z nowotworami głowy i szyi. Drugi zestaw obejmuje widma masowe dla sześciu typów nowotworów, których próbki dane są w postaci mikromacierzy. Rozdział opisuje szczegółowo pozyskanie i przygotowanie próbek

oraz pomiar widm masowych za pomocą spektrometrii MALDI-TOF. Dla pierwszego zestawu próbek pozyskano łącznie 73 225 widm, dla drugiego 50 905 widm. Biorąc pod uwagę, że każde widmo składa się z 100–200 tyś. punktów, należy docenić skalę badań eksperymentalnych.

W rozdziale piątym opisano metody wstępnego przetwarzania danych. Dane generowane przez spektrometry obciążone są błędami, które wynikają z niedoskonałego przygotowania próbki i kalibracji przyrządów oraz zakłócenia pomiaru. W pierwszym etapie wygładza się widmo i usuwa linię bazową (zmienny poziom intensywności bazowej). Do estymacji linii bazowej Autorka proponuje metodę adaptacyjną, która w odróżnieniu od metod standardowych, używa zmiennej szerokości okna estymacji. Linia bazowa dopasowywana do danych modelowana jest splajnami kubicznymi. W kolejnym etapie przetwarzania danych identyfikuje się obserwacje odstające. Autorka dokonuje przeglądu klasycznych metod detekcji obserwacji odstających, zauważając ich nieadekwatność do analizy widm masowych, z uwagi na asymetrię rozkładu danych. Ostatecznie rekomenduje metodę zaproponowaną dla rozkładów skośnych z ciężkimi ogonami, która wykrywa wartości odstające po obu stronach rozkładu. W dalszej części Autorka opisuje problem wyrównywania widm pozyskanych z tej samej próbki. Jako metodę wyrównywania widm stosuje algorytm oparty na szybkiej transformacie Fouriera. W podrozdziale 5.3 Autorka proponuje modelowanie widm masowych za pomocą mieszaniny rozkładów Gaussa. Każdy pik jest modelowany za pomocą jednego lub większej liczby komponentów gaussowskich. Prowadzi to do znacznej redukcji danych, liczba cech opisujących widmo zmniejsza się ok. dziesięciokrotnie. Aby uniknąć indywidualnego modelowania widma w każdym punkcie pomiarowym próbki biologicznej, Autorka proponuje wyznaczenie modelu średniego, dla całego zbioru widm jednej próbki. Model średni nie jest jednak zdefiniowany (patrz uwaga 6). Aby uzyskać reprezentacje poszczególnych widm, dokonuje się operacji „splotu" widma średniego z indywidualnymi widmami. Ta operacja nie jest w pracy dostatecznie zdefiniowana (patrz uwaga 6). W wyniku otrzymuje się nowe składowe widma, które wyrażają połączenie tych dwóch sygnałów. Autorka zauważa, że wiele składowych widma średniego wyrażonych komponentami gaussowskimi jest biologicznie nieistotnych. Słusznie proponuje usunięcie składowych o dużej wariancji i niewielkiej amplitudzie. Dokonuje detekcji tych składowych na podstawie histogramów współczynnika zmienności i amplitudy, które dekomponuje mieszaninami gaussowskimi. To pomysłowa metoda, chociaż mogłaby być bardziej szczegółowo opisana w zakresie wyboru punktów progowych na podstawie bayesowskiego kryterium informacyjnego. Zredukowany w ten sposób zbiór cech prowadzi do utraty informacji, ale głównie w końcowej części widma (m/z > 3000 Da), która ma mniejsze znaczenie w formowaniu biomarkerów.

Rozdział szósty opisuje budowę modelu klasyfikacyjnego i wykrywanie lokalnych wzorców molekularnych dla różnych typów nowotworów (drugi zestaw danych). Do oceny różnic w proteomie pomiędzy tkanką zdrową a nowotworową zastosowano miary wielkości efektu Cohena oraz Eta-kwadrat (niestety w pracy brak definicji tych miar). Wyniki analiz wskazują, że zmiany nowotworowe w nowotworach hormonozależnych mają charakter bardziej globalny i dotyczą większej liczby białek (ponad połowa cech może potencjalnie wskazywać na zmiany nowotworowe) niż w nowotworach spowodowanych czynnikami środowiskowymi. Podrozdział 6.2 omawia problem redukcji wymiarowości. Autorka opisuje szereg standardowych metod redukcji wymiaru takich jak analiza głównych składowych, ale orientuje się też w najnowszych rozwiązaniach, np. *uniform manifold approximation and projection* (UMAP). Wykorzystując tę ostatnią metodę, obrazuje na płaszczyźnie widma pozyskane z próbek różnych nowotworów. Wynik jest bardzo wartościowy: pozwala ocenić wizualnie złożoność biologiczną nowotworów, ich heterogeniczność i separowalność pomiędzy tkanką zdrową a nowotworową. Na tej podstawie można wnioskować o trudnościach w diagnozowaniu

pacjentów cierpiących na różne typy nowotworów. Słusznie zauważając, że metody redukcji wymiaru, operując w nowych przestrzeniach „gubią" informację biologiczną i tracą na interpretowalności, Autorka proponuje używanie ich jedynie do wizualizacji danych. Do budowy modelu klasyfikacyjnego rekomenduje metody filtracji cech omówione w rozdziale piątym (wykorzystujące mieszaninę gaussowską i usuwające składowe o dużej wariancji i małej amplitudzie). Metody te pozwoliły uzyskać wyniki porównywalne z metodami redukcji wymiarowości opisywanymi w tym rozdziale, redukując liczbę cech do kilkudziesięciu, co należy uznać za bardzo dobry rezultat.

W dalszej części rozdziału 6 opisano model klasyfikacyjny oparty na regresji logistycznej. Autorka zwraca uwagę na złożoność problemu klasyfikacji, która wynika z kilku czynników. Tkanka nowotworowa wykształca się z tkanki zdrowej w sposób „ciągły", co implikuje trudności z klasyfikacją przez patologa tkanki do kategorii zdrowa/nowotworowa. Duża heterogeniczność, złożoność procesów i mnogość zmian zachodzących w obu typach tkanek utrudnia tę klasyfikację. W badaniach eksperymentalnych Autorka rozważa dwa problemy modelowania bioróżnorodności: pomiędzy pacjentami cierpiącymi na ten sam typ nowotworu oraz różnorodności wewnątrztkankowej. Ten drugi problem wymaga użycia algorytmu grupowania danych. Autorka używa metody k-średnich w wersji odpornej na dane odstające i dobiera liczbę klastrów bazując na statystyce zaproponowanej przez Tibshiraniego (*gap statistics*) i kryterium Dunna. Wyniki grupowania wskazują większą heterogeniczność tkanki zdrowej niż nowotworowej. Podrozdział 6.6 omawia problematykę selekcji cech – kryteria selekcji cech oraz algorytmy *lasso* oraz *ridge*. Autorka proponuje własną metodę selekcji cech opartą na rankingu cech. W badaniach eksperymentalnych pokazuje jej przewagę nad metodami *lasso* i *ridge* – uzyskane sygnatury są znacznie krótsze przy podobnym poziomie błędu klasyfikacji.

W podrozdziale 6.7 Autorka optymalizuje model klasyfikacyjny. Pierwszym z problemów, który rozwiązuje jest problem niezbalansowania danych. Aby uniknąć tworzenia syntetycznych danych, proponuje bilansować klasy, losując próbki z powtórzeniami z mniej licznej klasy. Drugim problemem jest ustalenie progu w klasyfikatorze opartym na regresji logistycznej, decydującego o położeniu płaszczyzny decyzyjnej. Do znalezienia jego optymalnej wartości wykorzystuje krzywą ROC. Dyskutując konsekwencje nieprawidłowej klasyfikacji w rozpoznawaniu nowotworów, proponuje maksymalizować czułość klasyfikatora. Zoptymalizowany klasyfikator pozwolił uzyskać lepszą dokładność od klasycznego podejścia bez wyrównywania liczebności klas.

W rozdziale siódmym Autorka bada globalny profil proteomiczny raka. Zauważa, że choć różne typy nowotworów różnią się znacznie pod względem budowy molekularnej, to wiele z nich spowodowanych jest typowymi mutacjami. Mają więc wspólne cechy na poziomie molekularnym. Sensowne jest zatem podejście globalne, które polega na budowie modelu klasyfikacyjnego nie dla pojedynczych typów nowotworów, lecz dla wielu typów jednocześnie. Stosując ten sam schemat budowy klasyfikatora na podstawie widm masowych, co dla pojedynczych typów nowotworów, Autorka znajduje sygnaturę globalną dla drugiego zestawu danych i analizuje jej podobieństwo z sygnaturami poszczególnych typów nowotworów. Porównuje wyniki klasyfikacji w podejściu globalnym i lokalnym, trafnie konkludując, że podejście globalne jest generalnie mniej dokładne niż lokalne, ale dostarcza dodatkowych cennych informacji na temat sygnatury globalnej wyłonionej w procesie konstrukcji klasyfikatora.

Rozdział ósmy omawia budowę klasyfikatora dla pierwszego zestawu danych bazującego na widmach dla peptydów i lipidów. Autorka stosuje wypracowany schemat przetwarzania danych, ekstrakcji cech segmentacji tkanki na regiony zdrowe i nowotworowe, selekcji cech i klasyfikacji za pomocą regresji logistycznej. Dochodzi do ważnego wniosku, że domena peptydowa jest lepsza od lipidowej zarówno

w celach diagnostyczno-wizualnych jak i klasyfikacyjnych. Jednorodność obszarów lipidowych jest większa od peptydowych (większe skupiska), co skutkuje mniejszą dokładnością rozpoznawania obszarów zdrowych i nowotworowych. Autorka słusznie rekomenduje klasyfikator utworzony dla widm peptydów, który zapewnia wyższe wartości wszystkich przyjętych wskaźników oceny od klasyfikatora utworzonego dla lipidów.

W podsumowaniu pracy, rozdział dziewiąty, Autorka odwołuje się do jej tez i przekonująco udowadnia, że zostały one wykazane. Bibliografia pracy nie jest zbyt obszerna, ale zawiera wiele nowych pozycji ściśle związanych z tematyką rozprawy.

Sekwencja treści prezentowych w kolejnych rozdziałach jest właściwa: od informacji wstępnych dotyczących podłoża biologicznego i wykorzystywanej aparatury, poprzez opis pozyskiwania próbek biologicznych, przygotowania i przetwarzania danych, po budowę modeli klasyfikacyjnych i prezentację wyników. Praca napisana jest poprawnym językiem naukowym z właściwym słownictwem specjalistycznym i odpowiednią ścisłością sformułowań, choć nie jest wolna od błędów językowych (wymieniam te błędy w sekcji *Uwagi językowe, edytorskie i redakcyjne*). Układ redakcyjny rozprawy budzi jednak pewne zastrzeżenia. Praca sprawia wrażenie miejscami chaotycznej, np. problem redukcji wymiarowości omawiany jest niepotrzebnie w dwóch miejscach – w rozdziale piątym i podrozdziale 6.6, problem kroswalidacji omawiany jest w podrozdziale zatytułowanym *Inter-patient biodiversity*, wzory zapisane są niedbale, występuje tu wiele błędów, w tekście pojawiają się oznaczenia i pojęcia niezdefiniowane lub zdefiniowane w dalszej części pracy (szczegółowe uwagi w następnej sekcji).

Tezy rozprawy zostały wykazane. Znaczenie uzyskanych wyników dla rozwoju dyscypliny inżynieria biomedyczna oceniam bardzo wysoko w kontekście ich zastosowań do diagnostyki nowotworów i projektowania nowych leków zapobiegających nowotworom.

Uwagi krytyczne i polemiczne

1. Wzór (1) na stronie 21 jest błędnie zapisany. Druga para nawiasów okrągłych jest zbędna. Wzór (5) na stronie 49 jest błędnie zapisany. Brakuje znaku „–" przed wykładnikiem.

2. Opis metody adaptacyjnej korekcji linii bazowej zamieszczony w podrozdziale 5.1 jest mało czytelny. Brakuje szczegółowego algorytmu postępowania i graficznej wizualizacji metody. Nie jest jasne w jaki sposób do wykrywania trendu w oknie używa się współczynnika korelacji Pearsona, jaką rolę pełni 10% kwantyl sygnału oraz w jaki sposób dopasowuje się splajny kubiczne. Zamieszczone wyniki oceny proponowanej metody (CV) wymagają szerszego komentarza. Nie wiadomo dla jakich danych je uzyskano i jak wyglądał eksperyment badawczy.

3. W podrozdziale 5.2 Autorka opisała problem wykrywania obserwacji odstających. Jak sama zauważa, wartości odstające mogą reprezentować poprawne dane lub wskazywać na błędy w procesie akwizycji lub w aparaturze pomiarowej. Jak odróżnić te dwa przypadki? Jak wykorzystuje się informacje pozyskane w analizie danych odstających? Metody wykrywania obserwacji odstających opisane przez Autorkę wskażą najwyższe piki sygnału widma jako wartości odstające, a przecież są to najistotniejsze dane identyfikujące skład molekularny badanej próbki.

4. Na stronie 50 Autorka wprowadza pojęcie widma średniego reprezentującego cały zbiór danych. Nie opisuje jednak jak to widmo jest konstruowane. Widmo średnie ma kluczowe znaczenie, ponieważ służy do ekstrakcji cech reprezentujących indywidualne widma. Na podstawie tych cech działa system klasyfikujący. Sama procedura ekstrakcji cech też nie jest czytelna: „*In practice, the*

obtained GMM model was superimposed on each spectrum, and then the area under the curve was summed to estimate the amount of a specific peptide with a specific weight on the m/z axis." O jaką krzywą (*curve*) chodzi? Z czym sumowane jest pole powierzchni pod tą krzywą? Do jakiej wartości m/z to pole jest przypisywane, aby otrzymać nową cechę widma? Czy operację ekstrakcji cech można nazwać splotem modelu średniego z widmem?

5. W podrozdziale 6.1 do pomiaru zróżnicowania pomiędzy tkanką zdrową i nowotworową zastosowano miary wielkości efektu Cohena oraz Eta-kwadrat. Niestety nie podano definicji tych miar i nie uzasadniono ich użycia. Utrudnia to ocenę otrzymanych wyników.

6. Na stronie 69 zaprezentowano wyniki klasyfikacji dla różnych metod redukcji wymiarowości, nie wspominając o postaci klasyfikatora. Opis metody klasyfikacji używanej w badaniach powinien mieć miejsce przed prezentacją wyników.

7. Regresja logistyczna, której użyto do klasyfikacji danych jest metodą klasyfikacji liniowej. Czy Autorka jest pewna, że klasy reprezentujące tkankę zdrową i nowotworową są separowalne liniowo? W rozdziale 6.3 brakuje dyskusji tego zagadnienia. Brakuje też szerszego omówienia modelu regresji logistycznej: jej wielowymiarowej definicji, funkcji straty, estymacji parametrów, optymalizacji, wad i zalet.

8. Opis strategii uczenia modeli zamieszczony w podrozdziale 6.4. *Inter-patient biodiversity* na stronach 75-77 powinien znaleźć się w odrębnym podrozdziale. Nie odpowiada tematyce podrozdziału 6.4.

9. Podrozdział 6.6 na stronach 83-85 ponownie omawia problem redukcji wymiarowości i selekcji cech, który prezentowany był w rozdziale piątym. Całe to zagadnienie powinno być opisane w rozdziale piątym.

10. Na stronie 84 Autorka błędnie podaje, że kryteria informacyjne (AIC, BIC) to metody selekcji cech.

11. Tabela 5 na stronie 86 jest nieczytelna: co oznaczają symbole K, dHart i bits?

12. Przyjęte kryterium ustalania progu odcięcia w metodzie selekcji cech opisanej na stronach 88-89 (*knee method*) jest niejasne. Sprawia wrażenie subiektywnego.

13. Na stronie 93 Autorka zapowiada modyfikację krzywej ROC (wykres NPV w funkcji 1 – PPV) i selekcję progu odcięcia w regresji logistycznej na jej podstawie. Cel tej modyfikacji jest niejasny. Ponadto wykres pokazany na rysunku 24 prezentuje niezmodyfikowaną krzywą ROC.

14. W podrozdziałach 6.6 i 6.7 Autorka używa wskaźników jakości klasyfikacji, które definiuje dopiero w podrozdziale 6.8. Definicji niektórych wskaźników brakuje, np. PPV i NPV. Ważona dokładność (*weighted accuracy*) została błędnie zdefiniowana wzorem (26).

Uwagi językowe, edytorskie i redakcyjne

- W pracy występuje wiele skrótów, których nie objaśniono w miejscu ich pierwszego użycia lub wcale nie objaśniono, np. MSI, MALDI MSI na stronie 1 i 2, FFPE na stronie 11, DESI na stronie 18, PPV I NPV na stronie 93.
- Strona 4 – wyodrębnienie tylko jednego podrozdziału w rozdziale 7 nie wygląda dobrze.

- Strona 16 – brakuje „to” po „due”, zbędne „s” w 8 linii od dołu.
- Strona 20 – niezrozumiałe wtrącenie ("Thomson Joseph John").
- Na stronie 50 ostatnie zdanie drugiego akapitu jest powtórzone jako pierwsze zdanie trzeciego akapitu.
- Wzór (6) na stronie 50 – zmienne w nawiasie powinny być rozdzielone przecinkami.
- Strona 54 – symbole zmiennych w tekście pisane są prostą czcionka zamiast kursywą.
- Strona 55 – komponenty gaussowskie na rysunku 12 są słabo widoczne. To samo na rysunku 20.
- Rysunek 16 na stronie 60 – pomylono kolejność pików na jednej z osi.
- Strona 62 – „featires” zamiast „features”.
- Strona 68 – „abudance” i „abudndance” zamiast „abundance”.
- Strona 68 – brak opisu osi x na rysunku 20.
- Strona 71 i 87 – powinno być “kernel function” zamiast “nuclear function”.
- Strona 72 – przy definicjach β_0 i β_1 brakuje znaków „–” po prawej stronie znaków równości.
- Strona 81 – „samle” zamiast „sample”, „meanurement” zamiast „measurement”.
- Strona 85 – niepotrzebne powtórzenia przy definicji czynnika Bayesa.
- Strona 88 – „feture” zamiast „feature”.
- Strona 94 – „a real class” zamiast „a predicted class”.
- Strona 95 – “story matrix” zamiast “confusion matrix”.

Wniosek końcowy

Zakres tematyczny rozprawy doktorskiej mgr inż. Katarzyny Frątczak i osiągnięte w niej oryginalne wyniki w zakresie rozpoznawania nowotworów na podstawie obrazowania molekularnego metodą spektrometrii mas lokują tę rozprawę w obszarze inżynierii biomedycznej. Uważam, że rozprawa stanowi oryginalne rozwiązanie problemu naukowego i wskazuje na wysoki poziom wiedzy Autorki z dyscypliny inżynieria biomedyczna, a także na umiejętność samodzielnego prowadzenia przez nią badań naukowych. Pomimo zamieszczonych powyżej uwag krytycznych moja generalna opinia o pracy jest pozytywna.

Stwierdzam, że opiniowana rozprawa doktorska spełnia wymogi ustawy z 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce. Wnoszę o dopuszczenie mgr inż. Katarzyny Frątczak do publicznej obrony pracy doktorskiej i wyróżnienie rozprawy doktorskiej.