Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science
Institute of Informatics

Doctor of Biomedical Engineering Dissertation

Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data

mgr inż. Wojciech Sikora

Supervisor: Prof dr hab. inż. Joanna Polańska

mgr inż. Wojciech Sikora

Zastosowanie technik uczenia maszynowego do kompleksowej analizy danych obrazowania molekularnego MALDI-TOF w badaniach nad rakiem

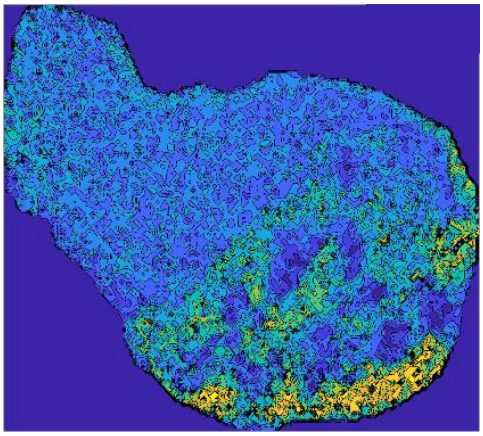
Streszczenie

Przedmiotem pracy doktorskiej jest analiza danych otrzymanych za pomocą obrazowania spektrometrią mas próbek pobranych od pacjentów z nowotworem głowy i szyi. Celem pracy jest zaproponowanie kompletnego schematu przetwarzania danych, począwszy od surowych danych otrzymanych z obrazowania tkanek, a skończywszy na modelu zdolnym do przypisania nowych obserwacji do jednej z wielu klas. W ramach pracy postawiono trzy hipotezy. Pierwsza hipoteza dotyczy metody identyfikacji pików w spektrach masowych. Hipoteza twierdzi, że identyfikacja pików może być skutecznie przeprowadzona przy pomocy modelu spektrum utworzonego poprzez podzielenie spektrum na części, a następnie dopasowaniu do tych części, modelu mieszanin normalnych. Druga hipoteza dotyczy metody usuwania redundancji w danych oraz zmniejszania wymiarowości danych. Druga hipoteza twierdzi, że informacja o przestrzennej dystrybucji danych dostępna dzięki obrazowaniu spektrometrią mas, może być wykorzystana do skutecznej eliminacji redundancji i znacznego zmniejszenia wymiarowości danych, przy jednoczesnym zachowaniu jakości danych. Trzecia hipoteza dotyczy wnioskowania o ważności cech dla modeli trenowanych na heterogenicznych danych. Ostatnia hipoteza twierdzi, że identyfikacja najważniejszych cech w niejednorodnych danych jest możliwa i skuteczna dzięki wykorzystaniu wielu modeli jednostkowych.

Pierwsze rozdziały skupiają się na podstawowych zagadnieniach związanych z proteomiką i spektrometrią mas. Pierwszy rozdział pokrótce wyjaśnia dlaczego analiza danych z obrazowania spektrometrią mas jest istotna oraz wprowadza najważniejsze zagadnienia z nią związane. Zagadnienia te są rozwinięte w dalszych

częściach pracy. W pierwszym rozdziale poruszony jest temat wyzwań jakie analityk danych napotyka podczas pracy ze spektrami masowymi, a także omówiono jakie aktualnie metody procesowania takich danych są najczęściej wykorzystywane.

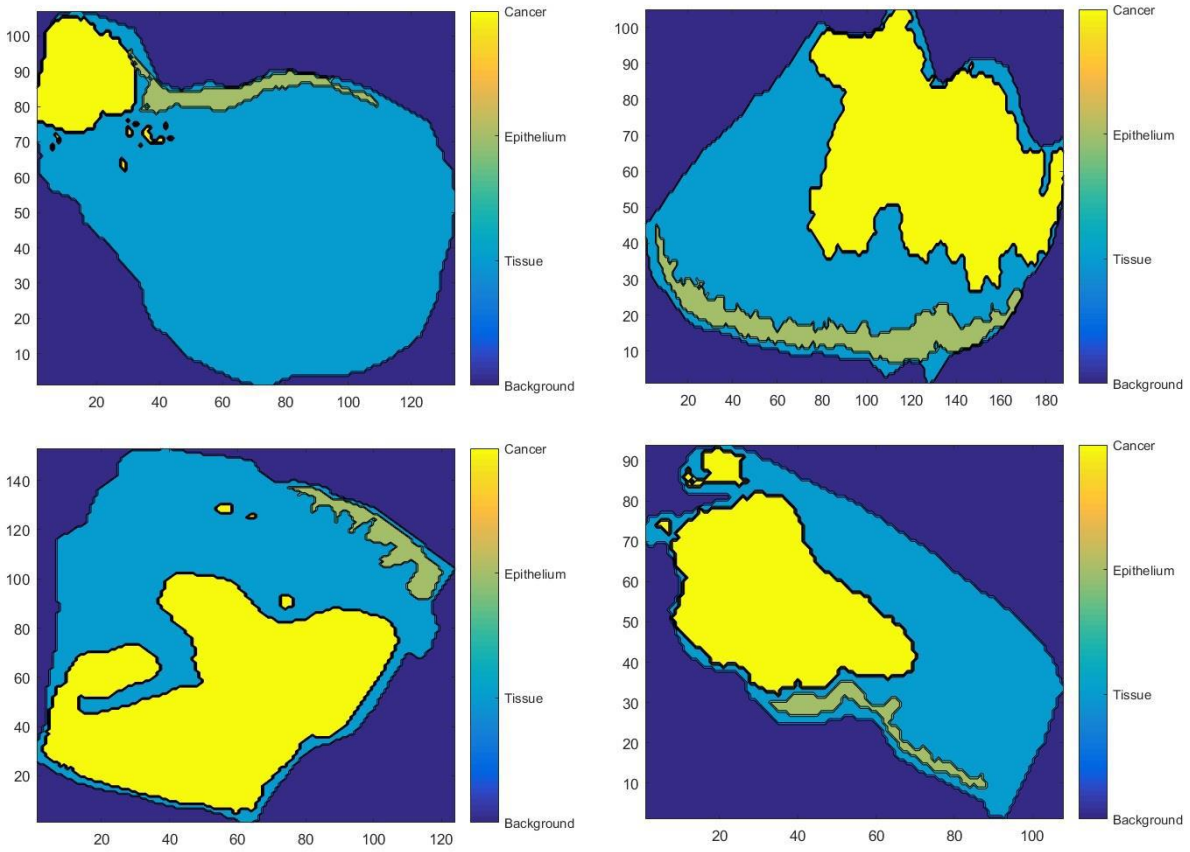
Drugi rozdział przedstawia z większą szczegółowością techniki pozyskiwania spektrów masowych oraz obrazowania tkanek z wykorzystaniem spektrometrii mas. W pierwszej kolejności przedstawiono ogólny schemat działania spektrometru oraz typowe parametry, które rozróżniają między sobą typy spektrometrów i mają znaczący wpływ na ich zastosowanie. Następnie wytłumaczone zostało czym dokładnie jest spektrum masowe, jak przebiega proces obrazowania za pomocą spektrometrii mas, a także jak wygląda wynik tego procesu (zobacz Rys. 1).



Rysunek 1: Przykładowy obraz uzyskany dzięki obrazowaniu spektrometrią mas.

Dalej opisane są najistotniejsze, z punktu widzenia obrazowania, metody jonizacji, pierwszego etapu spektrometrii mas. Opisano trzy metody jonizacji, które są powszechnie wykorzystywane przy obrazowaniu biologicznych próbek. Te metody to desorpcja i jonizacja przez elektrorozpylanie (Desorption electrospray ionization – DESI), desorpcja laserowa wspomagana matrycą (Matrix-assisted laser desorption ionisation – MALDI) oraz spektrometria mas jonów wtórnych (Secondary ion mass sepctrometry – SIMS). Rozdział zawiera opis zasady działania tych metod wraz z wizualizacjami oraz porównanie parametrów jakie oferują. W rozdziale są również wspomniane inne metody, które najczęściej są zmodyfikowanymi wersjami wymienionych, ale nie są jeszcze powszechnie wykorzystywane do badań. Dalej rozdział opisuje następne etapy spektrometrii mas czyli, separacja i detekcja jonów. W szczególności opisano metodę separacji jonów na podstawie czasu przelotu jonów (time of flight - TOF), która to metoda została wykorzystana do otrzymania przetwarzanych w ramach pracy danych.

Rozdział trzeci skupia się na przedstawieniu danych przetwarzanych w ramach pracy. Na początku opisano proces pobrania próbek od pacjentów z rakiem szyi i głowy, referując krok po kroku działania jakie zostały wykonane przez specjalistę, przyjęte parametry, a także dokładny model i ustawienia spektrometru. Pobrane próbki z zaznaczonymi regionami nowotworu, nabłonka i normalnej tkanki pokazuje rysunek 2.



Rysunek 2: Pobrane próbki z zaznaczonymi regionami nowotworu, nabłonka i normalnej tkanki.

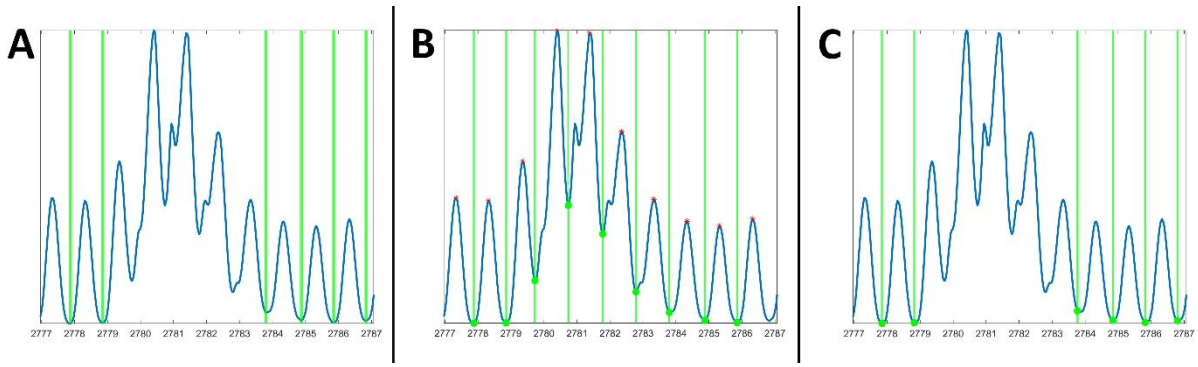
W rozdziale opisano również pierwsze kroki podjęte w celu przygotowania surowych danych do kluczowej analizy, czyli identyfikacji pików. W pierwszej kolejności opisano proces usuwania przesunięcia linii bazowej. Podano również inne powszechnie stosowane metody korekcji linii bazowej ja podano inne typowe metody jej usuwania jak na przykład metody oparte na transformacie falkowej lub metody wykorzystujące funkcje sklejane. Dalej podano sposób normalizacji danych. Jest to standardowa metoda TIC (total ion current/count), która polega na policzeniu sumarycznej intensywności, każdego punktu pomiarowego i podzieleniu wartości w każdym punkcie przez tę wartość. Ostatnie podane działanie to wyrównanie spektrów za pomocą szybkiej transformaty Fouriera. W rozdziale podano również odniesienia do

najbardziej przydatnych i ogólnie dostępnych narzędzi umożliwiających lub pomagających przy wykonywaniu tych kroków.

Czwarty jest wprowadzeniem do najważniejszego etapu przetwarzania danych ze spektrometrii mas, czyli identyfikacji pików. W tym rozdziale przedstawiono aktualny stan wiedzy na temat detekcji pików, w szczególności dla danych związanych z proteomiką. W pierwszej kolejności opisano proces agregacji wszystkich dostępnych spektrów masowych do pojedynczego zagregowanego spektrum, na którym wykonywane są dalsze operacje. Przedstawiono i porównano trzy różne możliwości uzyskania zagregowanego spektrum. Na zagregowanym spektrum porównano różne metody identyfikacji pików. Identyfikacje pików wykonano na początku za pomocą prostej metody wybierania pików na polegającej na filtrowaniu pików poniżej wybranej wartości progowej dla stosunku sygnału do szumu dla intensywności. Szum zdefiniowano jako odchylenie bezwzględne w lokalnym sąsiedztwie. Jest to prosta i bardzo często wykorzystywana metoda, która jest zdolna wykryć tylko piki o intensywności powyżej poziomu szumu. Wyniki zastosowania tej metody pokazano przyjmując różne parametry dla szerokości okna, dla którego określa się lokalny szum oraz różnych wartości progowych stosunku sygnału do szumu. Wyniki okazały się niesatysfakcjonujące. W następnej kolejności wypróbowano metodę opartą na transformacie falkowej. Jest to metoda zaliczająca się do kategorii metod modelowania piku, które oprócz intensywności, wykorzystują także kształt piku do identyfikacji. W pierwszym kroku tej metody sygnał jest transformowany do przestrzeni falkowej, następnie w tej przestrzeni następuje identyfikacja „linii grzbietu", czyli miejsc o wysokim pokryciu falki ze spektrum. W pracy podano matematyczny opis transformaty falkowej, a także dokładny opis zasady działania algorytmu. Uwzględniono metody identyfikacji linii grzbietowych oraz wizualizację kolejnych etapów algorytmu, a także rezultat identyfikacji pików tą metodą.

Piąty rozdział przedstawia docelową metodę identyfikacji pików zaproponowaną w ramach pracy jako jeden z etapów powtarzania danych w ramach przedstawianego schematu działania. Jest to metoda, której celem jest stworzenie modelu całego spektrum jako mieszaniny rozkładów normalnych. Metoda zaproponowana w podanych publikacjach, polega na podzieleniu sygnału na mniejsze części, a następnie dopasowaniu do każdej z nich mieszaniny rozkładów normalnych. Na początku rozdziału wyjaśniono motywację tego wyboru oraz argumenty przemawiające za wyższością takiego podejścia w stosunku do innych metod. Wynika ona natury spektrów masowych uzyskanych za pomocą spektrometrów wykorzystujących czas przelotu do separacji

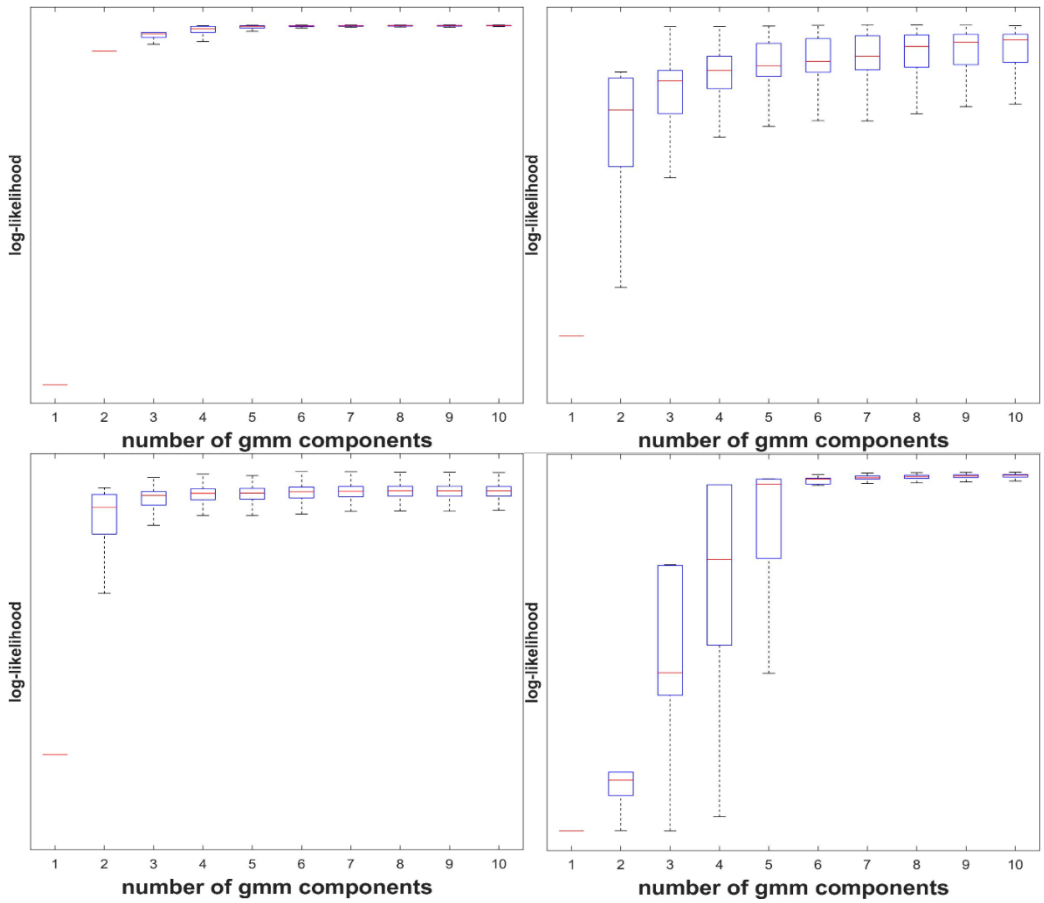
jonów oraz zastosowanej metody jonizacji. W dalszej części opisano badania jakie zostały wykonane w celu usprawnienia oryginalnej metody podziału spektrum na mniejsze części. Na początku opisano oryginalną metodę polegającą, w uproszczeniu, na identyfikacji „prawdziwych" pików i wykorzystaniu ich jako miejsc podziału. Następnie przeprowadzono rozważania nad metodą oceny w jaki sposób ocenić jakość podziału spektrum na części oraz ustanowiono dwie zasady, którymi należy się kierować przy podziale. Pierwsza zasada mówi, że części powinny być możliwie jak najmniejsze i jeśli to możliwe zawierać pojedynczy pik. Druga zasada mówi, że zgrupowania nakładających się na siebie pików nie powinny być rozdzielone i powinny się znaleźć w jednej części. Dalej przeprowadzono eksperymenty z dwoma nowymi metodami podziału. Pierwsza metoda, podobna do oryginalnej, polega na detekcji pików za pomocą transformacji falkowej, a następnie podzieleniu sygnału w miejscach pomiędzy znalezionymi pikami. Druga metoda skupia się od razu na znalezieniu optymalnych punktów podziału przez poszukiwanie lokalnego minimum. Metoda ta dodatkowo określa, na podstawie lokalnego sąsiedztwa, wartość progową, poniżej której lokalne minimum jest odrzucane i nie uwzględniane na liście miejsc do podziału sygnału. Obie metody oraz manualny podział zostały porównane na rzeczywistym przykładzie (zobacz Rysunek 3) i ocenione z wykorzystaniem zdefiniowanych kryteriów. Ostatecznie, metoda poszukiwania lokalnego minimum okazała się lepsza. W pracy zawarto także pseudokod tej metody oraz wizualizację wyników działania każdej z metod.



Rysunek 3: Porównanie metod poszukiwania punktów podziału spektrum. Podział manualny (A). Podział z wykorzystaniem transformaty falkowej (B). Podział przez poszukiwanie lokalnych minimum (C).

Dalsza część rozdziału skupia się na dopasowywaniu modeli mieszanin normalnych do przygotowanych części spektrum. Ponieważ analityczne wyznaczenie optymalnych parametrów jest niemożliwe, dopasowywanie mieszanin rozkładów normalnych jest wykonane z wykorzystaniem iteracyjnego algorytmu expectation-maximization (EM).

W tekście pracy zawarto zarówno słowny, jak i matematyczny opis tego algorytmu. Jest również pokazana wizualizacja procesu dopasowywania mieszaniny rozkładów normalnych do losowo wybranej części spektrum dla kolejnych iteracji algorytmu. Do przeprowadzenia badań wykorzystano własną implementację algorytmu, której pseudokod również zawarto w pracy. Ostatnie podrozdziały opisują proces wyznaczania parametru *k*, czyli liczby elementów mieszaniny, którą dopasowujemy do danej części spektrum. Jakość dopasowania jest określana przez obliczenie wiarygodności modeli oraz, na jej podstawie, kryterium informacyjnego bayes'a, aby wprowadzić karę za skomplikowanie modelu i uniknąć nadmiernego dopasowania. Ze względu na stochastyczną naturę algorytmu EM, dla każdej wartości parametru *k* dopasowywanie mieszaniny jest powtarzane wielokrotnie. Narysowane wykresy pudełkowe (zobacz Rysunek 4) obrazują jak zmienność wiarygodności modeli maleje wraz ze wzrostem liczby elementów. Po dopasowaniu mieszanin rozkładów normalnych do każdej części spektrum uzyskano model spektrum składający się z 9454 elementów.



Rysunek 4: Wykresy pudełkowe dla 4 losowo wybranych części spektrum.

Szósty rozdział pracy jest poświęcony inżynierii cech, której celem jest redukcja wymiarowości danych, a także usunięcie redundancji. Kluczowym aspektem tej części

jest porównywanie przestrzennej dystrybucji elementów modelu spektrum na fizycznych próbkach. W pierwszej kolejności opisano cele i zagrożenia związane z redukcją wymiarowości danych. Następnie przedstawiono strategie usuwania szumu z danych. Opiera się ona na wykorzystaniu parametrów dystrybucji gaussowskiej, a konkretnie parametru σ związanego z kształtem piku oraz parametru λ określającego intensywność piku. Zbadana została dystrybucja całej populacji wartości tych parametrów. Następnie dopasowano mieszaninę rozkładów normalnych i wyznaczono wartość krytyczną wykorzystaną do odfiltrowania elementów modelu sklasyfikowanych jako szum. W wyniku redukcji szumu, liczba elementów modelu została zredukowana z 9454 do 2884.

Redukcja liczby elementów modelu jest następnie kontynuowana poprzez porównanie przestrzennej dystrybucji pobliskich elementów (cech) na fizycznych próbkach. Podjęcie takiego działania jest zmotywowane faktem, że modelowanie rozkładami normalnymi może modelować pojedynczy pik jako wiele elementów modelu. Jest to spowodowane przez niedoskonałości procesu korekty linii bazowej, a także faktem, że typowy kształt piku nie jest idealnym rozkładem normalnym. Ze względu na niedokładności pomiarowe i drobne różnice podczas ruchu jonów w polu magnetycznym, piki w spektrum masowym mają nieznacznie skrzywiony kształt, charakteryzujący się spłaszczeniem po prawej stronie. Dla każdej części spektrum pozostałe elementy modelu są porównywane między sobą. Dystrybucja przestrzenna elementów jest porównywana za pomocą testu Peacock'a, który jest rozszerzeniem testu Kołmogorowa-Smirnowa do dwóch wymiarów. W pierwszej kolejności szczegółowo opisano zasadę działania testu Kołmogorowa-Smirnowa, wraz z podaniem przykładu i wizualizacją tej statystyki testowej przy porównaniu dwóch sztucznie wygenerowanych dystrybuant. Następnie przedstawiona jest zasada działania testu Peacocka, również z wizualizacją oraz przykładem działania na rzeczywistych losowo wybranych elementach modelu spektrum. Ponieważ, dla testu Peacocka nie istnieją poziomy istotności, wartości krytyczne zostały wyznaczone przez numeryczne symulacje. Ponieważ każda próbka ma inny rozmiar, wyznaczono wartość krytyczną dla każdej próbki. Kilka tysięcy wartości statystyki testowej Peacocka obliczono dla każdej próbki, a następnie na podstawie dystrybucji tych wartości określono wartości progowe dla testów wskazujących na identyczną, podobną lub całkowicie różną dystrybucję przestrzenną. Wartość *p* dla testów jednostronnych dla każdej próbki połączono w pojedynczą wartość *p*, korzystając z metody Fishera. W wyniku tych operacji liczba elementów modelu zmniejszyła się z 2884 do 2392.

Ostatni podrozdział jest poświęcony detekcji obwiedni izotopowych. Obwiednia izotopowa jest ekspresją konkretnego peptydu, w którego składzie chemicznym występują różne izotopy atomów, powodując różnice w masie, a co za tym idzie różne wartości stosunku masy do ładunku (m/z). Obwiednie izotopowe utrudniają analizę widma masowego i korzystne jest przedstawienie ich w postaci pojedynczej cechy w miejscu występowania dominującego piku. Najczęściej różnica w masie atomowej pomiędzy kolejnymi izotopami wynosi 1 Da. Ponadto piki należące do jednej obwiedni izotopowej powinny mieć podobny kształt, a także ich dystrybucja przestrzenna powinna być taka sama. Wykorzystując te informacje stworzony został algorytm wyszukujący obwiednie izotopowe. W pierwszej kolejności sprawdzana jest odległość pomiędzy elementami (różnica pomiędzy wartościami parametru μ elementów modelu spektrum), jeżeli odległość znajduję się w numerycznie wyznaczonym zakresie poprawnych wartości odpowiadających różnicy w masie równej 1 Da, badany jest kształt pików (stosunek wartości σ). Jeżeli oba warunki są spełnione, porównywana jest przestrzenna dystrybucja pików. Piki spełniające wszystkie trzy kryteria są uznane za część obwiedni izotopowej, a poszukiwanie jest kontynuowane. Zidentyfikowane obwiednie izotopowe są zastąpione pojedynczym elementem modelu w miejscu dominującego izotopu, a wartość powstałej w ten sposób cechy jest obliczana na podstawie sumy wartości otrzymanych z wszystkich elementów obwiedni izotopowej.

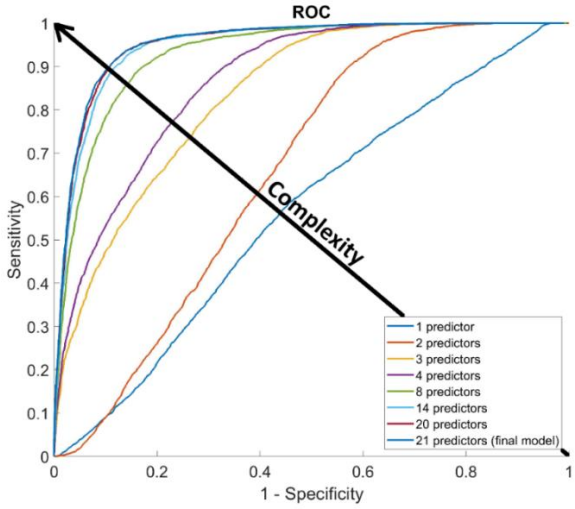
Na końcu rozdziału podsumowano wyniki działania inżynierii cech. Ostatecznie model spektrum składający się z 9454 elementów zredukowano do zbioru 888 cech, potwierdzając słuszność drugiej postawionej w pracy hipotezy. Liczba cech na poziomie kilkuset jest zgodna z oczekiwaniami na podstawie spodziewanej liczby molekuł w tego typu próbkach o intensywności dostatecznej do ich wykrycia.

Rozdział 7 jest poświęcony zastosowaniu metod statystycznych oraz uczenia maszynowego w celu nauczenia, na przetworzonych danych, klasyfikatorów zdolnych do podjęcia decyzji o przynależności nowej obserwacji do konkretnej klasy. W pierwszej kolejności podano jakie metody najczęściej są wykorzystywane w publikacjach o takiej tematyce. Następnie poruszony jest temat podziału danych na zbiory do kroswalidacji. Z całego zbioru danych najpierw wydzielono zbiór walidacyjny o wielkości 10% całego zbioru, a następnie dla każdego modelu jednostkowego podział na zbiór treningowy i testowy został wykonany za pomocą próbkowania przy zachowaniu balansu pomiędzy klasami (stratified sampling).

Dalej podano wytłumaczenie oraz wzory wszystkich miar obliczanych w celu oceny jakości klasyfikacji. Są to podstawowe miary obliczane na podstawie macierzy błędów takie jak dokładność, precyzja, czułość, swoistość, wartość predykcyjna ujemna, miara

F1. Wyjaśniono również, dlaczego nie należy wnioskować o jakości modelu na podstawie pojedynczej miary. Następnie szczegółowo wytłumaczono metody porównywania jakości modeli predykcyjnych, z wykorzystaniem krzywych oraz pól pod krzywymi ROC, wykresów balansu pomiędzy precyzją i czułością (precision - sensitivity), a także dodatnią oraz ujemną wartością predykcyjną (PPV - NPV).

Następny podrozdział opisuje metodę uczenia klasyfikatorów opartą na wielomianowej regresji logistycznej. Metoda ta polega przeprowadzeniu regresji logistycznej po kolei, wykorzystując każdą z cech z finalnego zbioru cech jako zmienną objaśniającą. Cecha o najlepszej wiarygodności zostaje wybrana jako najistotniejsza cecha i tym samym jako pierwszy element modelu. Następnie regresja jest powtarzana po kolei dla każdej z pozostałych cech, tym razem z uwzględnieniem wcześniej wybranych cech. Ponownie cecha, która razem z wcześniej wybranymi cechami uzyskała najlepszą wiarygodność jest wybierana jako kolejny element modelu. Proces jest powtarzany aż do momentu, w którym na podstawie czynniku Bayesa, podjęta jest decyzja, że wzrost wiarygodności modelu nie kompensuje dostatecznie wzrostu poziomu skomplikowania modelu związanego z dodaniem kolejnego elementu. W pracy pokazano jak jakość modelu zmienia się wraz z dodawaniem kolejnych cech do listy zmiennych objaśniających regresji, zarówno dla krzywej ROC (Rysunek 5), jak wykresu Precyzja-Czułość. Trenowanie klasyfikatora w opisany sposób zostało powtórzone wielokrotnie dla losowych podziałów na zbiory treningowe i testowe, a wyniki przedstawiono w postaci tabeli ze średnimi wartościami miar jakości wraz z 95% przedziałami ufności. Dodatkowo opisano proces możliwej dalszej optymalizacji w celu uzyskania lepszego kompromisu pomiędzy konkretnymi miarami jakości klasyfikacji, jak na przykład dodatniej oraz ujemnej wartości predykcyjnej poprzez maksymalizację współczynnika Youdena, a także wyniki klasyfikacji po korekcie.



Rysunek 5: Wpływ poziomu skomplikowania modelu na jakość klasyfikacji.

Drugą grupę klasyfikatorów stanowią proste sieci neuronowe. Na takich samych podziałach na zbiory treningowe i testowe nauczono klasyfikatory z wykorzystaniem w pełni połączonych sieci neuronowych z dwoma ukrytymi warstwami o liczbie węzłów równej liczbie cech. Wyniki zostały pokazane w sposób analogiczny do metody opartej na regresji logistycznej. W porównaniu do bardziej rozbudowanej metody iteracyjnego przeprowadzania regresji logistycznej, proste sieci neuronowe okazały się gorsze, lecz również charakteryzują się bardzo dobrymi miarami jakości klasyfikacji. Potwierdza to wysoką jakość procesu przetwarzania danych.

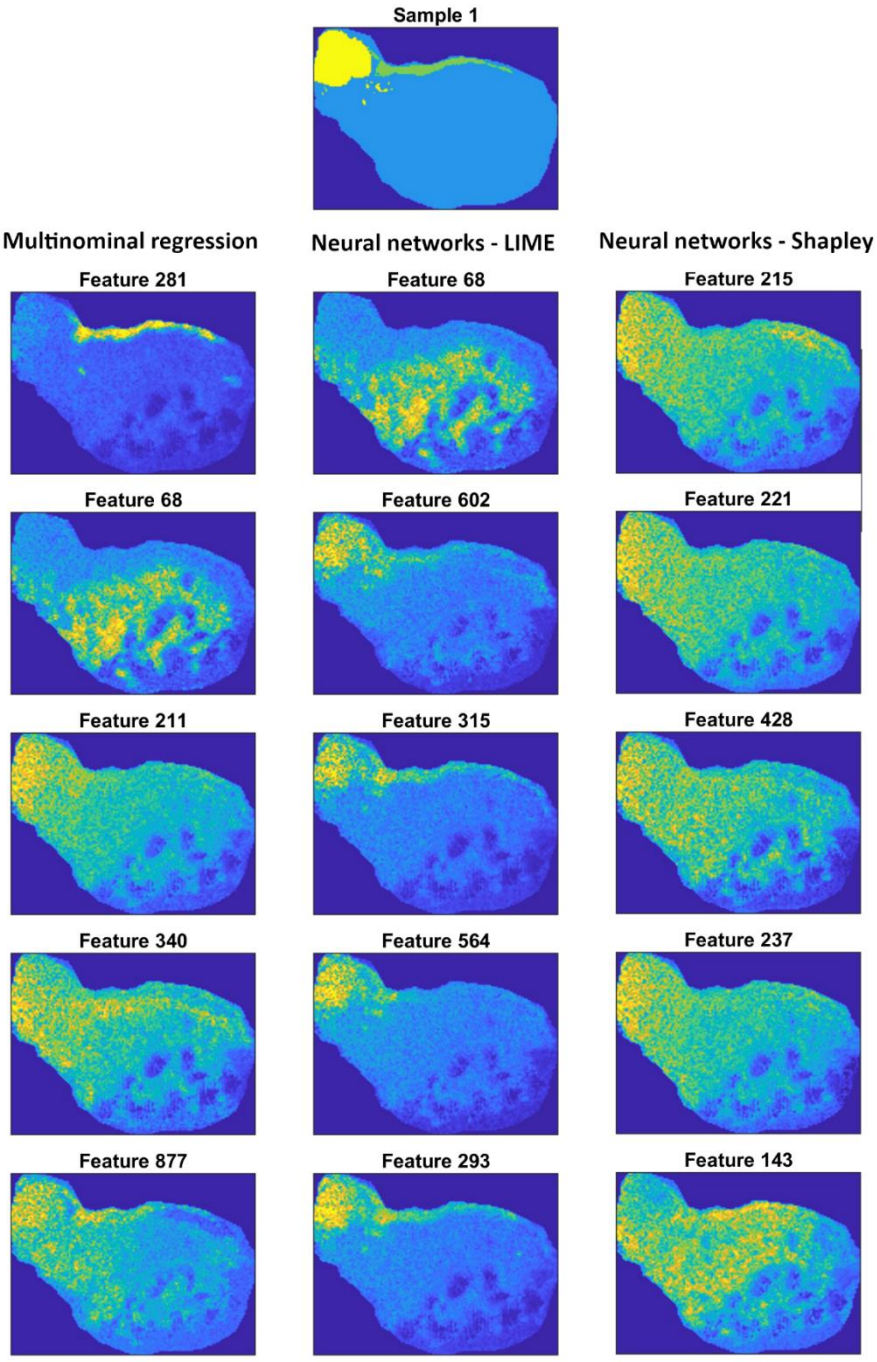
Ostatnie podrozdziały pracy są poświęcone zadaniu określenia ważności cech, a tym samym weryfikacji ostatniej hipotezy postawionej w ramach pracy. Zaproponowano metodę określenia ważności cechy dla obu grup klasyfikatorów. Klasyfikatory nauczone z wykorzystaniem algorytmu opartego na regresji logistycznej są uporządkowanymi listami cech. Ogólna ważność cechy została obliczona jako średni wynik danej cechy z wszystkich modeli jednostkowych, gdzie wynik jest równy $\frac{1}{x}$, gdzie x to pozycja cechy na liście.

W celu określenia ważności cech dla sieci neuronowych wykorzystano dwie metody interpretacji. Pierwszą metodą jest LIME (local interpretable model-agnostic explenation). Jest to metoda interpretacji, którą można wykorzystać dla każdego modelu. Stara się ona objaśnić, które cechy konkretnej obserwacji mają największy wpływ na wynik klasyfikacji poprzez zaburzanie wartości cech np. poprzez losowanie nowej wartości z normalnej dystrybucji o parametrach obliczonych na podstawie całej populacji wartości tej cechy. Następnie nowo powstała obserwacja jest klasyfikowana za pomocą objaśnianego modelu. Po zebraniu dostatecznie wielkiego zbioru danych jest

na nich uczony interpretowalny klasyfikator np. drzewa decyzyjne, dzięki czemu można wnioskować o ważności cech. Objaśniając w ten sposób wiele tysięcy obserwacji, wyciągnięto wnioski o globalnej ważności cech.

Drugą z metod są wartości Shapleya. Jest to metoda ewaluacji ważności cech dla modeli „czarnych skrzynek", opierająca się na założeniu, że każda cecha ma swój udział w ostatecznym wyniku klasyfikacji oraz, że można zbadać ten udział poprzez weryfikację jak usunięcie danej cechy wpływa na wynik. Takie wartości oblicza się dla wszystkich możliwych podzbiorów cech, przez co czas obliczeń dokładnych wartości Shapleya rośnie wykładniczo wraz z liczbą cech. Z tego względu większość implementacji w tym ta wykorzystana w ramach pracy oblicza przybliżone wartości Shapleya na ograniczonej liczbie podzbiorów. Kolejnym czynnikiem negatywnie wpływającym na jakość wyników jest fakt, że cechy nie da w prosty sposób usunąć i trzeba taką sytuację zasymulować poprzez losowanie wartości z populacji i uśrednianie wyników. Rezultat określania ważności cech przedstawiono przez pokazanie pięciu najistotniejszych cech według każdej z metod (Rysunek 6).

Ostatni rozdział zawiera rozważania na temat przeprowadzonych badań, wniosków z nich płynących oraz planów na kontynuowanie badań.



Rysunek 5: Wizualizacja najważniejszych cech dla próbki numer 1.

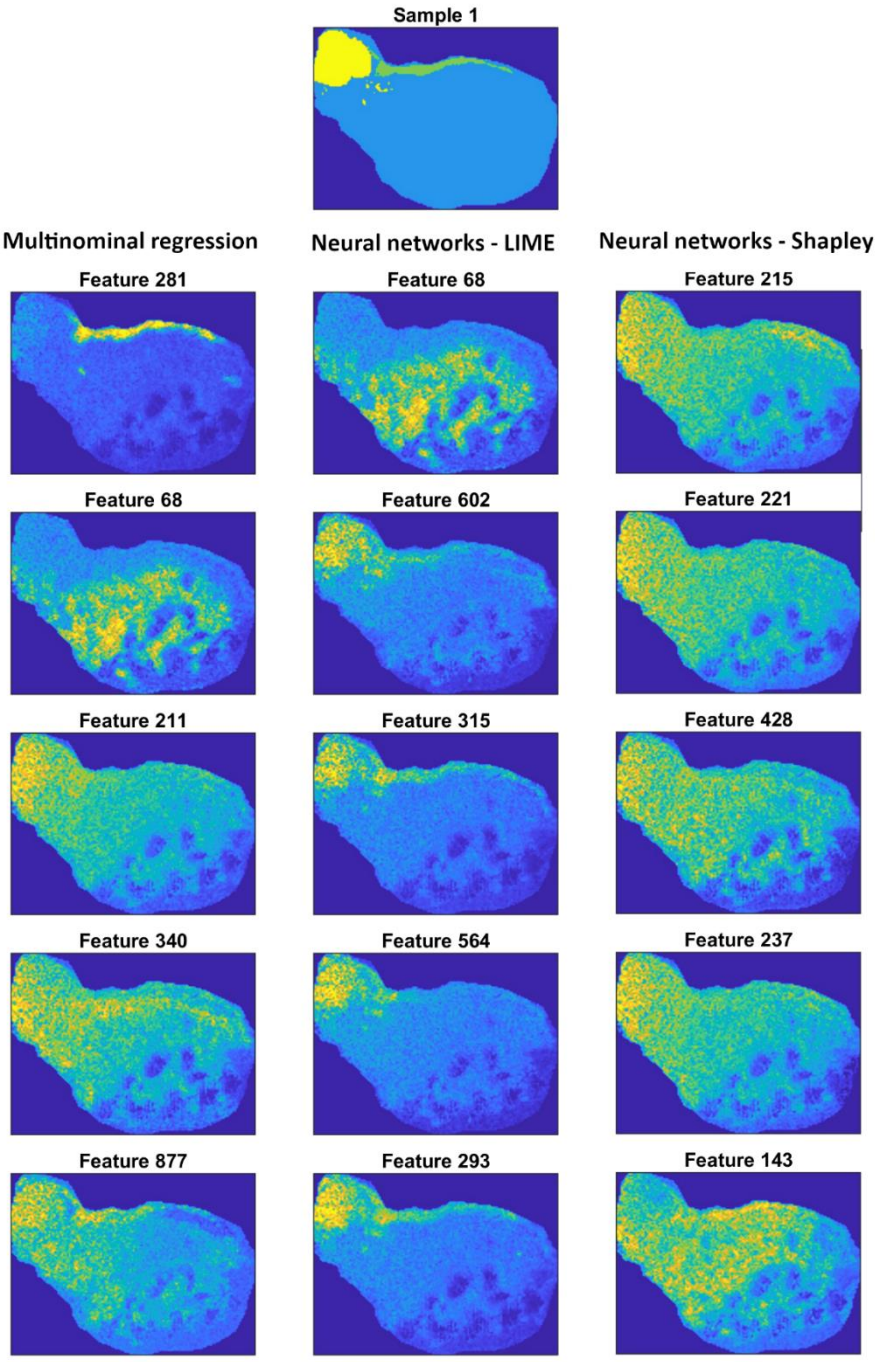


Figure 5: Images for top five features for sample number 1.