POLITECHNIKA WARSZAWSKA

Wydział Mechatroniki

ROZPRAWA DOKTORSKA

mgr inż. Zofia Lorenc

System do klasyfikacji obiektów warstwowych wykorzystujący techniki spektralne VIS

Promotor
prof. dr inż. Leszek Sałbut

Promotor pomocniczy
dr inż. Sławomir Paśko

Warszawa 2020

Podziękowania

Doktorant jest odpowiedzialny za swój statek, nie za fale.
Parafraza słów Abrahama Lincolna

Z początku płynęłam wpław. Otaczający ludzie podawali mi kolejne składowe tratwy. Zbudowałam ją. Przemieszczałam się już szybciej i bezpieczniej, jednak nie zawsze w odpowiednim kierunku. Nieraz odwracaliśmy z Profesorem łódź o 180°. Chwilami sztorm życia próbował roztrzaskać łajbę. Jednak zawsze przychodził ktoś z pomocą, dawał narzędzie do załatania wyrwy, wypychał z mielizny, kawałek podholował… Mijał czas i łódź, dzięki tym wszystkim wspaniałym ludziom spotkanym po drodze i mojej determinacji, zamieniła się w solidny statek. Fale życia nadal w niego uderzały, ale on nieustraszenie mknął do celu. Teraz, jako trójmasztowy żaglowiec, dostojnie zacumował w doktorskim porcie.

Dziękuję każdemu, kto przyczynił się do zbudowania fregaty, jak również wszystkim wskazującym słuszny kierunek podczas tej skomplikowanej podróży, a także ludziom, którzy uśmiechem dodawali mi otuchy.

Streszczenie

Gwałtowny wzrost ilości, rozmiarów, jakości czy dokładności gromadzonych danych związany jest z szybkim rozwojem technik inżynierskich i informatycznych stosowanych w dziedzinie rejestracji sygnałów. W dzisiejszym świecie coraz częściej wykonuje się różnego rodzaju analizy: obrazu, dźwięku czy sygnału spektralnego, wykorzystując do tego celu algorytmy klasyfikacji. Nieraz obiekty opisywane są przez dziesiątki, setki, a nawet tysiące atrybutów, co może generować problemy. Zdarza się, że wielowymiarowość sygnałów nie idzie w parze z liczbą poddanych analizie próbek. W takich przypadkach konieczna jest selekcja cech. Znanych jest wiele algorytmów stworzonych w tym celu, jednak w połączeniu z typowymi metodami klasyfikacyjnymi, w przypadku mało licznej próby, wyniki mogą okazać się niesatysfakcjonujące.

W niniejszej rozprawie zostało podjęte wyzwanie częściowego rozwiązania ww. problemu, poprzez zaprojektowanie systemu do klasyfikacji obiektów warstwowych wykorzystującego techniki spektralne promieniowania z zakresu widzialnego — VIS, wraz z opracowaniem autorskiej metody parametryzacji z użyciem aproksymacji wielomianowej (PAW) oraz metody doboru i redukcji widma (DRW). Praca wykazuje potencjał stosowania zakresu VIS w klasyfikacji obiektów metodami spektralnymi. Prezentowany system łączy w sobie punktowość pomiarów spektroskopowych z pewnego rodzaju uogólnieniem interpretacji sygnału występującym w obrazowaniu spektralnym. Opracowany system charakteryzuje się dużą elastycznością dostosowania do konkretnego przypadku. Dowodzi tego opis dwóch przeprowadzonych eksperymentów: klasyfikacja skorup jaj kurzych pod względem występowania patogenu u zwierzęcia, od którego pochodziła badana skorupa oraz klasyfikacja miodów pod względem pochodzenia botanicznego.

Praca wpisuje się dziedzinę spektroskopii. Uzyskując wyniki porównywalne z technikami wykorzystywanymi dotychczas, wykazuje możliwość zastosowania znacznie węższego widma promieniowania w celach klasyfikacyjnych obiektów warstwowych. Wnosi ona również nowe podejście do redukcji wymiarowości sygnałów spektralnych zaprezentowane w metodach PAW i DRW. Są one w szczególności dedykowane do zastosowań, w których liczy się interpretacja fizyczna zarejestrowanych widm. Cecha ta może zostać wykorzystana w celu adaptacji systemu do konkretnego zastosowania, co przekłada się na korzyści ekonomiczne. Dzięki zaimplementowanym algorytmom system może być stosowany w przypadku, gdy metody typu sztuczne

sieci neuronowe czy maszyna wektorów nośnych nie są w stanie wygenerować prawidłowych modeli ze względu na niewystarczającą liczbę próbek.

Słowa kluczowe: *techniki spektralne, spektroskopia VIS, klasyfikacja, redukcja wymiarowości, redukcja cech.*

Abstract

The rapid increase of the amount, size, quality or accuracy of the collected data is associated with the fast development of engineering and IT techniques used in the field of signal acquisition. Nowadays it is becoming increasingly common to perform various types of analysis, i.e. image, sound or spectral signal analysis, using classification algorithms. Objects are sometimes described by tens, hundreds and even thousands of attributes, which can generate problems. The multidimensionality of signals does not always go in parallel with the number of samples analysed. In such cases, feature selection is necessary. There are many well known algorithms created for this purpose, but in combination with typical classification methods, in the case of a small amount of samples, the results may be unsatisfactory.

This dissertation deals with the challenge of partially solving the abovementioned problems by developing a system for the classification of layered objects using the spectral techniques of visible range of electromagnetic spectrum (VIS), together with the development of a new parameterization method based on the polynomial approximation method (PM) and spectrum selection and reduction method (SSR). The presented system combines the spectroscopic spot measurements with a way of generalisation of signal interpretation occurring in spectral imaging. The developed system is characterised by high flexibility and adaptability to a specific use case, which was proved by a detailed description of two experiments: chicken egg shell classification in terms of a pathogen presence in the hen that laid it and honey classification in terms of its botanical origin.

The thesis is part of the field of spectroscopy indicating the possibility of using the VIS radiation range alone for the classification purposes of layered objects. Moreover, the dissertation brings a new approach to reduction of the dimensionality of spectral signals presented in the PM and SSR methods. It is particularly dedicated to applications where physical interpretation of registered spectra counts. This feature can be used to adapt the system to a specific application which can generate economic benefits. Thanks to implemented algorithms the system can be used in case where neural networks or support vector machine are not able to generate proper models due to an insufficient number of samples.

Keywords: *spectral techniques, VIS spectroscopy, classification, dimensionality reduction, feature reduction.*

Spis treści

1. Wstęp . . . 13

1.1. Struktura pracy . . . 13

1.2. Wprowadzenie . . . 14

1.3. Cel pracy . . . 17

2. Przegląd literatury . . . 18

2.1. Techniki spektralne wykorzystujące promieniowanie elektromagnetyczne . . . 18

2.1.1. Podstawy fizyczne . . . 18

2.1.2. Rejestracja widma fal elektromagnetycznych . . . 21

2.2. Przetwarzanie danych intensywnościowych oraz klasyfikacja . . . 24

2.2.1. Ogólne podejście do procesu klasyfikacji . . . 24

2.2.2. Wstępne przetwarzanie danych . . . 26

2.2.2.1. Eliminacja wpływu charakterystyki podłoża i źródła światła . . . 27

2.2.2.2. Operacja filtracji . . . 30

2.2.3. Selekcja cech . . . 33

2.2.3.1. Redukcja wymiarowości - generacja nowych cech . . . 33

2.2.3.2. Redukcja wymiarowości - selekcja widmowa . . . 37

2.2.4. Klasyfikacja w uczeniu statystycznym . . . 38

Klasyfikatory statystyczne i minimalnoodległościowe . . . 40

Drzewo decyzyjne . . . 41

Sztuczne sieci neuronowe . . . 43

Maszyna wektorów nośnych . . . 47

2.2.5. Miary oceny jakości klasyfikacji . . . 49

2.3. Podsumowanie rozdziału . . . 51

3. Metoda parametryzacji i metoda redukcji sygnału . . . 52

3.1. Idea działania proponowanych metod . . . 52

3.2. Analiza . . . 56

3.2.1. Metoda parametryzacji z użyciem aproksymacji wielomianowej – PAW . . . 57

3.2.2. Metoda metoda doboru i redukcji widma – DRW . . . 61

3.3. Wpływ niedokładności i rozdzielczości pomiaru urządzenia na działanie metod 63
3.4. Porównanie metod selekcji cech 67
3.5. Podsumowanie rozdziału 69

4. Adaptacyjność systemu wykorzystującego metody PAW i DRW na przykładach 72

4.1. System 72
4.2. Przykłady zastosowań systemu 72
Wybór technik porównywanych 74
4.2.1. Klasyfikacja skorup jaj kurzych ze względu na efekt działania Mycoplasma Synoviae 76
Wstęp 76
Materiał badawczy 77
Układ optyczny 78
Pomiary 80
Analiza i wyniki 80
Podsumowanie i wnioski 85
4.2.2. Klasyfikacja miodów ze względu na pochodzenie botaniczne 89
Wstęp 89
Materiał badawczy 90
Układ optyczny 90
Pomiary 92
Anliza i wyniki 92
Podsumowanie i wnioski 96
4.3. Podsumowanie rozdziału 98

5. Podsumowanie rozprawy 99

5.1. Wnioski, podsumowanie rozprawy oraz kierunki dalszych prac 99
5.2. Elementy nowości w pracy 101

Spis rysunków 102

Spis tabel 106

Bibliografia 108

Wykaz oznaczeń stosowanych w rozprawie

acc	dokładność klasyfikacji drzewa decyzyjnego z kroswalidacją (ang. accuracy) [1]
$acc\,r.$	współczynnik dokładności klasyfikatora (ang. accuracy rate)
CVA	analiza wariancji kanonicznej (ang. canonical variate analysis)
HCA	hierarchiczna analiza klastrów (ang. hierarchical cluster analysis)
ANN	sztuczne sieci neuronowe (ang. artificial neural networks)
CART	drzewa klasyfikacyjne i regresyjne (ang. classification and regression trees)
CHAID	automatyczny rejestrator interakcji za pomocą chi-kwadrat (ang. chi-squared automatic interaction detector)
DT	drzewo decyzyjne (ang. decision tree)
DRW	autorska metoda doboru i redukcji widma
FIR	promieniowanie elektromagnetyczne z zakresu dalekiej podczerwieni (ang. far infrared)
IR	promieniowanie elektromagnetyczne z zakresu podczerwieni (ang. infrared)
IQR	zakres międzykwartylowy (ang. interquartile range)
k-NN	metoda najbliższego sąsiada (ang. k nearest neighbours)
LDA	liniowa analiza dyskryminacyjna (ang. linear discriminant analysis)
MR	rezonans magnetyczny (ang. magnetic resonance)
MS	patogen *Mycoplasma Synoviae*
MSC	multyplikatywna korekcja rozproszenia (ang. multiplicative scatter correction)
NIPALS	nieliniowa iteracyjna metoda najmniejszych kwadratów (ang. nonlinear iterative partial least squares)
NIR	promieniowanie elektromagnetyczne z zakresu bliskiej podczerwieni (ang. near infrared)
NMR	magnetyczny rezonans jądrowy (ang. nuclear magnetic resonance)
N-W	Norris-Williams
PARAFAC	równoległa analiza czynnikowa (ang. parallel factor analysis)
PAW	autorska metoda parametryzacji z użyciem aproksymacji wielomianowej
PC	składowa główna uzyskana z metody PCA (ang. principal component)
PC1/2/3	pierwsza, druga i trzecia składowa główna uzyskana z metody PCA
PCA	analiza składowych głównych (ang. principal component analysis)
PLS	metoda cząstkowych najmniejszych kwadratów (ang. partial least squares)

PLS-DA analiza dyskryminacyjna cząstkowych najmniejszych kwadratów (ang. partial least square prediction — discriminant analysis)

QUEST szybkie nieobciążane wydajne drzewo statystyczne (ang. quick unbiased efficient statistical tree)

RBF sieć oparta na radialnych funkcjach bazowych (ang. radial basis function)

SFS/SBS sekwencyjna selekcja postępująca/wsteczna (ang. sequential forward/backward selection)

SFFS ruchoma sekwencyjna selekcja postępująca (ang. sequential floating forward selection)

S-G Savitskiy-Golay

SIMCA proste modelowanie analogii klas (ang. simple modeling of class analogy)

SNR stosunek sygnału do szumu (ang. signal-to-noise ratio)

SVM maszyna wektorów nośnych (ang. support vector machine)

UV promieniowanie elektromagnetyczne z zakresu ultrafioletu (ang. ultraviolet)

VIS promieniowanie elektromagnetyczne z zakresu widzialnego (ang. visible)

WT analiza falkowa (ang. wavelet transform)

1. Wstęp

1.1. Struktura pracy

Niniejsza praca składa sie z pięciu rozdziałów.

W pierwszym przedstawiono ideę rozprawy, zdefiniowano cel pracy i wymieniono zadania, które pozwoliły osiągnąć wyszczególniony cel. Zostały wprowadzone i pokrótce scharakteryzowane sformułowania należące do dziedziny technik spektralnych takie jak spektroskopia, spektrometria i obrazowanie spektralne. We wstępie wskazano również dostrzeżone przez autorkę problemy w kontekście specyficznych przypadków klasyfikacji.

Drugi rozdział poświęcony jest ogólnej prezentacji obecnego stanu wiedzy z zakresu technik spektralnych wykorzystujących promieniowanie elektromagnetyczne i z zakresu przetwarzania danych intensywnościowych, w kontekście szeroko rozumianego procesu klasyfikacji obiektów. W rozdziale zarysowano fizyczne podwaliny procesów umożliwiających wykonanie spektroskopii ze szczególnym uwzględnieniem promieniowania VIS. Opisano zjawisko barwy. Omówiona została systematyka pomiarów spektralnych ze względu na rozmiar powierzchni, z której następuje akwizycja danych (pomiary punktowe i polowe). Scharakteryzowano systemy wielokanałowe (pomiary wielo-, super- i hiperspektralne) przedstawiając istniejące techniczne rozwiązania. W drugiej części rozdziału znajduje się ideologiczny opis procesu klasyfikacji próbek fizycznych z omówieniem jego poszczególnych etapów. Przedstawione zagadnienia opisywane są w kontekście pomiarów spektralnych. Na podstawie przeglądu literatury wybrano techniki selekcji cech i klasyfikacji, które zostały szczegółowo opisane, a następnie wykorzystane w części eksperymentalnej.

Dalsza część rozprawy poświęcona jest zagadnieniom ściśle związanym z prezentowanym systemem do klasyfikacji obiektów warstwowych wykorzystującym techniki spektralne VIS.

Rozdział trzeci zaznajamia czytelnika z ogólną ideą i technicznymi szczegółami opracowanych metod parametryzacji i redukcji, będących kluczowym elementem prezentowanego systemu. Przedstawiono w nim również analizę błędów pomiarowych oraz wpływ stosowania

różnych urządzeń na wynik końcowy klasyfikacji. Rozdział kończy zbiorcze porównanie ideowe opracowanego podejścia z metodami powszechnie wykorzystywanymi do selekcji cech sygnału w dziedzinie przetwarzania sygnałów wraz z podsumowaniem.

W rozdziale czwartym zaprezentowano system do klasyfikacji obiektów warstwowych wykorzystujący techniki spektralne VIS i pokazano jego adaptacyjność na przykładzie dwóch szczegółowo opisanych eksperymentów. Poszczególne etapy badań zawierają odnośniki do fragmentów niniejszej rozprawy. Rozdział kończy podsumowanie cech proponowanego systemu.

Rozprawę doktorską zwieńcza podsumowanie, zawierające wnioski płynące z zawartej w niej treści oraz wyszczególnione elementy nowości pracy.

1.2. Wprowadzenie

Początków spektroskopii optycznej można szukać w badaniach Isaaca Newtona (klasyczny eksperyment z pryzmatem wyjaśniający zjawisko dyspersji w szkle). Widma różnych substancji wykorzystywali w swoich pracach Max Planck [2], Albert Einstein [3], czy Niels Bohr [4] już w pierwszych latach XX w.

Metody pomiarowe, których celem jest rejestracja natężenia promieniowania w funkcji długości fali, powszechnie nazywa się technikami spektralnymi. Przyjęło się wykorzystywać zwrot „spektroskopia" do pomiarów rejestrujących promieniowanie elektromagnetyczne oddziałujące na atomy i cząsteczki testowanego materiału, np. spektroskopia ultrafioletu (UV), spektroskopia promieniowania z zakresu widzialnego (VIS), spektroskopia podczerwieni (IR), spektroskopia ramanowska, absorpcyjna spektroskopia promieniowania X, czy spektroskopia rezonansu magnetycznego (NMR). Badania wykorzystujące inne zakresy widmowe zazwyczaj określa się jako „spektrometria", np. spektrometria masowa czy elektronowa. Czasem jednak autorzy stosują to nazewnictwo wymiennie. Obie techniki, zarówno spektrometria, jak i spektroskopia, dają wiele możliwości rozwoju prac badawczych w chemii, fizyce, biologii, medycynie czy inżynierii. Innym sformułowaniem występującym w omawianej dziedzinie jest „obrazowanie spektralne". Ten termin w szczególności odnosi się do miernictwa geologicznego i dziedzin pokrewnych, jak również meteorologii, kryminalistyki, czy dziedzictwa kulturowego.

W spektroskopii UV, UV-VIS, VIS-IR, IR i pokrewnych rejestrowane widmo jest sumą odpowiedzi poszczególnych związków chemicznych na wzbudzenie wywołane padającą wiązką fal elektromagnetycznych. Czasem sygnały te nakładają się na siebie. Wielokrotnie celem

tego typu badań jest pomiar stężenia i identyfikacja pojedynczych związków chemicznych lub ich grup funkcyjnych. Im wyższa złożoność próbki, tym wyniki badania są trudniejsze do interpretacji. Podstawą tego typu analiz są przede wszystkim widma absorpcyjne, będące efektem pracy układów światła przechodzącego. Najczęściej wykorzystywaną aparaturą do tych pomiarów są spektrofotometry — względnie duże (ok. 100 x70 x40 cm), złożone urządzenia, zawierające między innymi często więcej niż jedno źródło światła oraz komorę przeznaczoną na próbkę. Konieczność umiejscowienia mierzonego obiektu w niewielkiej kubaturze komory, ogranicza możliwość wykorzystania tych technik. Ponadto, przed pomiarem wykorzystującym zarówno zakres UV jak i IR, w wielu przypadkach należy w odpowiedni, nieraz czasochłonny i drogi sposób przygotować próbkę. Często niezbędna jest wstępna obróbka fizyko-chemiczna wykorzystująca procesy rozcieńczania, rozpuszczania, wysuszania, roztarcia, stworzenia zawiesiny olejowej, czy sprasowanej pastylki [5]. Czasem należy wykonać pomiary widma suchych pozostałości z wyciągu wodnego, w innym przypadku zastosować rozpuszczalnik organiczny [6]. W wielu dziedzinach jak np. badania in vivo, kryminalistyka czy dziedzictwo kulturowe, ingerencja w próbkę jest wysoko niezalecana.

Inaczej wygląda sytuacja w obrazowaniu spektralnym. W przeważającej liczbie przypadków akwizycja danych następuje przy użyciu układu światła odbitego. W pomiarach geodezyjnych, czy w rolnictwie, próbką jest pewien obszar powierzchni Ziemi. W takim przypadku pojęcie przygotowania obiektu badań nie istnieje. Do obrazowania wykorzystuje się kamery multi- lub hiperspektralne, których rozdzielczość widmowa jest rzędu kilku, czy nawet kilkudziesięciu nanometrów, a obszarami pomiarowymi mogą być powierzchnie o boku od metra do kilku kilometrów [7, 8]. Zdarza się, że rejestrowane widmo zawiera luki w sygnale, spowodowane odseparowaniem zakresów widmowych poszczególnych detektorów, będących częściami składowymi urządzenia pomiarowego. Obrazowanie spektralne może być również wykorzystywane w pomiarach małych powierzchni, np. rzędu kilku milimetrów. W takim przypadku rozdzielczość spektralna jest często wyższa, np. 1 nm [9, 10].

Szybki rozwój technologii pomiarów spektralnych umożliwiający rejestrowanie coraz szerszych zakresów widmowych, z coraz większą rozdzielczością generuje konieczność korzystania z szybszych i wydajniejszych algorytmów analizy danych, lub zmiany procedur ich obróbki. Przy tak wysoko-wymiarowych pomiarach powszechnie wykorzystywane modele klasyfikacyjne wymagają do prawidłowego działania licznej próby zbioru uczącego. W wielu

przypadkach dostęp do nowych próbek próbek jest utrudniony, a czasem zupełnie nie możliwy do realizacji.

W przypadku korzystania z systemów klasyfikacyjnych bazujących na PCA – analiza składowych głównych, ANN – sztuczne sieci neuronowe, SVM – maszyna wektorów nośnych lub analogicznych, wynikiem jest wartość liczbowa niewskazująca zakresu widmowego, dla którego badany materiał cechuje się własnościami umożliwiającymi przydzielenie do klas. Używając takiego systemu nie jest możliwe dostosowanie go do konkretnego zastosowania. Może się to wiązać z niepotrzebnie dużymi kosztami np. wykorzystaniem spektroskopii UV-VIS w przypadku, gdy zadane badanie może być z powodzeniem przeprowadzone tylko na ograniczonym zakresie VIS. W takim przypadku wynikowy system zawierałby drogie elementy optyczne przystosowane do zakresu UV, a użytkownik nie byłby tego świadomy.

Kolejną problematyczną kwestią jest wskazanie ujednolicenia grubości badanych próbek mierzonych w świetle przechodzącym. Kiedy algorytmy obróbki danych uwzględniają poziomy intensywności zarejestrowanych sygnałów konieczna jest ich wzmożona kontrola, ponieważ łatwo można wprowadzić do klasyfikatora zafałszowaną informację informującą o grubości nałożonej warstwy materiału badanego, a nie bezpośrednio o jego właściwościach fizyko-chemicznych.

Próba częściowego rozwiązania dostrzeżonych problematycznych zagadnień, niejasnych w kontekście konkretnych zastosowań, została podjęta w rozprawie. Do tego celu wykorzystano techniki spektralne VIS.

Przy wielu atutach pomiarów spektralnych bazujących na szerokich zakresach spektralnych (takich jak: UV – IR) należy pamiętać, że wykorzystując jedynie zakres promieniowania VIS bardzo rzadko istnieje możliwość identyfikacji pojedynczych związków chemicznych. Wynikowe widmo często nie będzie zawierać wąskich pików absorpcji charakterystycznych dla konkretnych substancji. Jednak klasyfikacja za pomocą technik spektralnych VIS również obiektów z pozoru nierozróżnialnych wykonywana jest z powodzeniem. Jest to możliwe dzięki procesom cyfrowej obróbki sygnału, takich jak zaproponowana w pracy metoda parametryzacji z użyciem aproksymacji wielomianowej (PAW) oraz metoda doboru i redukcji widma (DRW), przy jednoczesnym prawidłowo dobranym klasyfikatorze. W rozprawie opisano badanie, w którym dokładność klasyfikacji bazująca na 7-krotnej kroswalidacji dochodzi do 98%, a analogiczny parametr obliczony walidacją prostą osiąga 100%. Zaprezentowano również, jak

działałby system w przypadku użycia rejestratora o niższej rozdzielczości spektralnej, lub większych niepewnościach pomiarowych. Wykorzystanie spektroskopii szerszych widm: UV-VIS lub VIS-IR, jak również obrazowania spektralnego, umożliwiłoby uzyskanie analogicznych do prezentowanych wyników klasyfikacji, jednak koszt tych metod jest nieporównywalnie większy od nakładów jakie trzeba ponieść, stosując zaproponowany w rozprawie system. Należy przypuszczać, że również zwiększyłaby się ilość zarejestrowanych danych, powodując wymuszenie zwiększenia mocy obliczeniowej potrzebnej do uzyskania zadowalających rezultatów.

1.3. Cel pracy

Celem niniejszej rozprawy jest wskazanie potencjału zastosowania pomiaru widma promieniowania z zakresu VIS umożliwiającego klasyfikację obiektów warstwowych, przy wykorzystaniu autorskich metod parametryzacji i redukcji sygnału. Analizie procesu klasyfikacyjnego poddano obiekty warstwowe, przez co rozumie się obiekty, których trzeci wymiar będący grubością lub wysokością, jest mniej istotny w stosunku do dwóch pierwszych. Przykładami takich elementów są między innymi liście, wszelkiego rodzaju powłoki (np. lakier), skorupy jaj, rozmazy substancji na szkiełkach mikroskopowych, folie.

Ogólny cel pracy został sformułowany następująco:

Opracowanie systemu wykorzystującego pomiary widma promieniowania z zakresu widzianego wraz z implementacją metod parametryzacji i redukcji sygnału do klasyfikacji obiektów warstwowych.

Składają się na niego następujące cele szczegółowe:

- analiza sygnału widmowego ze szczególnym uwzględnieniem opracowanej metody parametryzacji z użyciem aproksymacji wielomianowej oraz metody doboru i redukcji widma;
- implementacja algorytmów technik uczenia statystycznego;
- budowa układów optycznych pracujących w świetle przechodzącym z wykorzystaniem spektrometru działającego w zakresie promieniowania widzialnego;
  - — układ przystosowany do pomiarów punktowych;
  - — układ z poziomym torem optycznym;
  - — układ z pionowym torem optycznym;
- wykonanie serii badań eksperymentalnych wykorzystujących opisany w rozprawie system.

2. Przegląd literatury

2.1. Techniki spektralne wykorzystujące promieniowanie elektromagnetyczne

2.1.1. Podstawy fizyczne

Światło, rozumiane jako promieniowanie fal elektromagnetycznych z zakresu widma widzialnego, opisywane jest, jak pozostałe fale elektromagnetyczne, równaniami Maxwella. Promieniowanie rozchodzi się zgodnie z kierunkiem wyznaczonym przez wektor Poyntinga, czyli wzdłuż kierunku przepływu energii. Analizując rozchodzenie się promieni na zasadach optyki geometrycznej, przyjmuje się takie uproszczenia zjawisk, które umożliwiają przyjęcie założeń liniowej propagacji światła. W tym podejściu nie uwzględnia się zjawisk takich jak dyfrakcja czy interferencja. Niemożliwa jest również analiza spektralna promieniowania, ponieważ całą wiązkę uznaje się za zbiór jednorodnych promieni o długości fali dążącej do zera.

Optyka falowa, dzięki swojemu bardziej złożonemu opisowi, uwzględnia długości fali promieniowania. Dzięki temu możliwy jest dokładny opis np. dyspersji czy dyfrakcji, bez których analiza spektralna nie miałaby racji bytu.

Rozpatrując wiązkę światła natrafiającą na nieciągłość ośrodka, w którym się rozchodzi, można zaobserwować różne zjawiska. Mogą zachodzić one równocześnie na rozdzielnych zakresach spektralnych lub selektywnie, gdy cały zakres ulega jednemu zjawisku. W przypadku, gdy promienie przedostaną się na drugą stronę obiektu zaburzającego rozchodzenie się wiązki w ośrodku, mówimy o transmisji lub załamaniu. Gdy po przejściu przez obiekt początkowa wartość intensywności danego zakresu spektralnego zostanie zmniejszona lub w ogóle niezarejestrowana, będzie oznaczać to, że promienie uległy pochłonięciu lub odbiciu. Analizując wymienione zjawiska w skali mikro można przyjąć, że każde z nich jest formą rozproszenia promieni na cząsteczkach.

Rozpatrując nieciągłość ośrodka jako zbiór cząsteczek, można przeprowadzić analizę propagacji energii w jednej z nich. Do tego celu wykorzystuje się kwantowy opis światła (wzór 2.1).

$$E = h\nu = \frac{hc}{\lambda} \tag{2.1}$$

gdzie: E — energia promieniowania;
h — stała Plancka: 6,62607 x 10^{-34}[J·s];
ν — częstotliwość fali wypromieniowanej;
c — prędkość rozchodzenia się światła w próżni;
λ — długość fali wypromieniowanej.

Energia promieniowania jest odwrotnie proporcjonalna do długości fali, co oznacza, że im bardziej światło zbliża się do zakresu UV, tym większą niesie ze sobą energię.

Zakładając pewne uproszczenia, do rozważań można przyjąć model, w którym elektrony znajdują się na powłokach o ściśle określonych poziomach energetycznych, a cząsteczkę traktuje się jako mały oscylator. W przypadku dostarczenia do cząsteczki kwantu energii o wartości równej różnicy poziomów energetycznych (częstotliwość rezonansowa), elektrony przedostają się na wyższy poziom energetyczny, następuje zjawisko absorpcji energii i cząsteczka ulega wzbudzeniu. Bezpośrednio po wprowadzeniu cząsteczki w ruch drgający następuje reemisja światła, która ma takie same właściwości falowe (np. ta sama długość fali) jak promienie pobudzające cząsteczkę. W taki submikroskopowy sposób można opisać makroskopowe zjawisko rozproszenia światła.

Ze zjawiskiem rozproszenia w sposób bezpośredni wiąże się pojęcie barwy. Może być ona wygenerowana na różne sposoby. Jednym z nich jest zjawisko rozproszenia Rayleigha będące często wytłumaczeniem powstania występującej w naturze barwy niebieskiej (intensywność I światła docierającego do obserwatora oddalonego o R, w wyniku rozproszenia Rayleigha przez jedną cząstkę o średnicy d, dla niespolaryzowanego światła o długości fali λ i intensywności światła padającego I_0 opisuje wzór 2.2, gdzie n – współczynnik załamania światła materiału cząstki, θ – kąt rozproszenia).

$$I = I_0 \frac{1 + cos^2\theta}{2R^2} \left(\frac{2\pi}{\lambda}\right)^4 \left(\frac{n^2 - 1}{n^2 + 2}\right)^2 \left(\frac{d}{2}\right)^6 \tag{2.2}$$

Reakcja oscylatora jest tym silniejsza, im bliższa częstotliwości rezonansowej jest wiązka pobudzająca. Jest to równoznaczne z faktem, iż światło fioletowe jest najsilniej rozpraszane „na boki" (poza kierunek propagacji), w następnej kolejności światło o barwie niebieskiej, zielonej, żółtej itd. Prawo Rayleya opisuje to zjawisko stwierdzeniem, iż intensywność światła rozproszonego (I) jest odwrotnie proporcjonalna do czwartej potęgi długości fali (λ). Światło rozpraszane jest w każdym kierunku (θ – kąt rozproszenia). W przypadku energii wypromieniowanej przez Słońce składowej fioletowej jest na tyle mało, że najbliższa częstotliwość ulegająca rozproszeniu Rayleigha odpowiada barwie niebieskiej. Temu zjawisku zawdzięczamy nie tylko kolor nieba, ale również np. niebieski kolor oczu [11]. Należy nadmienić, że opisywane zjawisko zachodzi na cząsteczkach o średnicach (d) mniejszych niż długość światła ($d < \sim \lambda/15$), co jest prawdziwe dla większości atomów i cząsteczek.

W przypadku, kiedy absorpcja cząsteczki nie jest wystarczająco silna i niemożliwa jest reemisja, następuje przepuszczenie energii z pasma rezonansowego. W miarę przedostawania się promieniowania wgłąb materiału następuje jego stopniowe pochłonięcie, rozumiane jako wytracenie energii. Przykładem tego zjawiska, zwanego selektywnym pochłanianiem, jest zielononiebieski odcień wody. Częstotliwość rezonansowa cząsteczki H_2O zawiera odcienie czerwonego z zakresu VIS, jednak absorpcja jest niewystarczająca aby nastąpiła reemisja tej barwy, więc promieniowanie o tym zakresie spektralnym zostaje przepuszczone i stopniowo pochłaniane do momentu całkowitego zaniku energii tych częstotliwości.

Nadawanie koloru materiałowi za pomocą cząsteczek pigmentów wykorzystuje również zjawisko selektywnego pochłaniania. W takim przypadku częstotliwości rezonansowe zawierają się w zakresie światła widzialnego. Dla pojedynczych atomów są to bardzo wąskie zakresy, jednak w przypadku ciał stałych i cieczy bliskość atomów powoduje poszerzenie poziomów energetycznych. Oznacza to, że częstotliwości rezonansowe obejmują szerokie zakresy częstotliwości, co przekłada się na pochłanianie szerokich zakresów spektralnych. W ten sposób odbieramy np. żółtą tkaninę, która pochłania w sposób selektywny barwę niebieską. Rozmiar cząstek jest czynnikiem silnie wpływającym na odbicie promieniowania, co zatem idzie, na kolor. Dlatego np. podczas badań spektralnych gleby zaleca się zmielenie próbki tak, aby osiągnęła formę drobnego miału ($< 10\ \mu\mathrm{m}$) [12].

Podczas opisu zjawiska pochłaniania warto uwzględnić również kwestie fosforescencji i fluorescencji. Możliwe jest, że cząsteczka nie wyemituje takiej samej energii, z jaką została wzbudzona, lecz mniejszą. Zjawisko to nazwane jest luminescencją. Cząsteczka w stanie

wzbudzonym może trwać około 0,1 s [13], następnie elektrony zmieniają położenie na niższą powłokę energetyczną, przy jednoczesnym wypromieniowaniu kwantu energii. Należy nadmienić, że wypromieniowany kwant charakteryzuje się niższą energią (co za tym idzie, dłuższą długością fali), niż ta, którą zaabsorbował, będąc pobudzanym. Część energii uległa rozproszeniu, np. podczas zderzeń z innymi cząsteczkami. Jeśli zjawisko to zaszło bezpośrednio (z poziomu tzw. singletowego), mówi się o fluorescencji. W przypadku, gdy elektron najpierw przedostanie się na tzw. poziom tripletowy i dopiero z niego, wyrównując energię do poziomu podstawowego, wypromieniuje kwant energii, mówi się o fosforescencji. Ze względu na trwałość stanu tripletowego, możliwe jest występowanie fosforescencji nawet do kilku minut, dni, a nawet lat po ustaniu promieniowania wzbudzającego [14].

Wzbudzając materiał energią z zakresu promieniowania VIS, następuje wzbudzanie energii oscylacyjnej. Oznacza to, że nie uwzględnia się energii wibracji atomów, energii związanych z rotacją cząsteczki wokół własnej osi, jak również energii elektronowej będącej sumą energii kinetycznej elektronów i potencjalnej między elektronami a jądrami atomów. W przypadku idealnym rejestrowane spektrum absorpcyjne cząstki wieloatomowej powinno przedstawiać wąskie piki odpowiadające poszczególnym składowym cząsteczki, jednak w praktyce obserwowane zazwyczaj zmiany intensywności zarejestrowanego widma są wolno zmienne w dziedzinie częstotliwości. Brak ostrych pików absorbcji ma swoje podłoże częściowo w efekcie Dopplera. Szerokość połówkowa profilu poszerzonej linii zależy od prędkości termicznej emitujących i absorbujących atomów oraz od rejestrowanej długości fali. Ponadto poszerzenie widma materiałów złożonych jest spowodowane jednoczesnym wzbudzaniem wielu oscylacji różnych atomów o nierównych różnicach energetycznych.

2.1.2. Rejestracja widma fal elektromagnetycznych

Do zagadnienia rejestracji widma fal elektromagnetycznych można podejść globalnie, makroskopowo — nie wnikając w strukturę materiałów, czy ruchów cząsteczkowych lub na bardzo wysokim poziomie szczegółowości — uwzględniając zjawiska wzbudzeń elektronowych, czy zmiany częstotliwości drgań w konkretnych zakresach widmowych.

Technika świetlna przede wszystkim koncentruje się na opisie zjawisk w rozumieniu makroskopowym. W tym podejściu nie uwzględnia się oddziaływań międzycząsteczkowych, większą wagę przykłada się do globalnego efektu często dostrzegalnego nieuzbrojonym okiem.

W przeważającej ilości przypadków do zadań dotyczących techniki świetlnej wykorzystuje się niekoherentne źródła światła opisywane za pomocą brył fotometrycznych.

Optyka, czy Fotonika opisują zjawiska świetlne na dużo wyższym poziomie szczegółowości. W tym podejściu, ze względu na zastosowania konieczny jest opis widma użytej wiązki światła z wysoką dokładnością. Bierze się również pod uwagę efekty falowe światła takie jak dyspersja czy interferencja. Jeszcze wyższy poziom szczegółowości opisu oddziaływania konkretnych fal na cząsteczki materiału wykorzystuje się w dziedzinie chemii. Możliwe tam jest określenie jakościowe i ilościowe składu chemicznego badanej substancji [15].

Techniki spektralne można podzielić ze względu na liczbę kanałów rejestrujących widmo w następujący sposób:

- pomiary wielospektralne — kilka, kilkanaście kanałów;
- pomiary superspektralne — kilkadziesiąt kanałów;
- pomiary hiperspektralne — kilkaset i więcej kanałów.

Podstawowym przedstawicielem pierwszej grupy jest rejestrator typu — kamera RGB [16]. W każdym pikselu obrazu zapisywana jest informacja o intensywności trzech kanałów — czerwonego, zielonego i niebieskiego. Rozwinięciem tego urządzenia są rejestratory super- i hiperspektralne, których zakresy widmowe przeważnie zawierają się od fal ultrafioletowych (UV), przez widzialne (VIS), do bliskiej (NIR), a nawet dalekiej podczerwieni (FIR).

Rejestracja poszczególnych kanałów jest możliwa w sytuacji, gdy analizowana wiązka światła ulegnie podziałowi na określone zakresy w dziedzinie widma. Taka filtracja może odbyć się po stronie detektora lub po stronie oświetlającego obiekt źródła światła. Przypadek filtracji po stronie obiektu można uzyskać następującymi sposobami [17]:

- Rejestracja nie więcej niż 6 kanałów — szerokopasmowe filtry absorpcyjne, często w połączeniu z filtrem Bayera [18, 19]. Jest to jedno z tańszych dostępnych rozwiązań, charakteryzuje się dużą niezawodnością, jednak niemożliwe jest uzyskanie wysokich rozdzielczości widmowych za jego pomocą;
- Rejestracja nie więcej niż 20 kanałów — wąskopasmowe filtry interferencyjne umieszczone na obrotowym kole przed rejestratorem [20]. Jest to często wykorzystywane rozwiązanie, jednak jedną z jego głównych wad jest zależność transmisji filtra od kąta padania wiązki światła;

- Rejestracja 30 kanałów i więcej — filtry przestrajalne (akustooptyczne i ciekłokrystaliczne filtry przestrajalne) [21]. Rozwiązanie charakteryzuje się bardzo wysoką rozdzielczością widmową, jednak wadą są problemy geometryczne z wiązką. Należy do grupy drogich rejestratorów.

Filtracja po stronie rejestratora odbywa się w samym urządzeniu. W spektroskopii funkcję detektora (mogącego być częścią składową spektrofotometru) najczęściej spełnia spektrometr. Można wyróżnić dwa najczęściej spotykane typy tych urządzeń: spektrometr siatkowy oraz pryzmatyczny, jak również spektrometr Fabry–Perot. W omówionych wcześniej układach podział na określone zakresy w dziedzinie widma następował dzięki odpowiednio dobranym filtrom, w przypadku spektrometrów funkcję tę pełni odpowiednio siatka dyfrakcyjna, pryzmat lub etalon w specyficznej konfiguracji. Rozszczepienie światła następuje dzięki zjawisku dyfrakcji na odbiciowej siatce dyfrakcyjnej, dyspersji widmowej w pryzmacie lub filtracji selektywnej interferometru Fabry–Perota [22].

Omawiając systematykę technik spektralnych należy uwzględnić również podział na pomiary punktowe oraz polowe. Obrazowanie, czyli rejestrowanie widma powierzchni wykorzystywane jest w takich zagadnieniach jak pomiary kolorymetryczne, charakterystyki źródeł światła i wyświetlaczy, czy szeroki wachlarz rozwiązań związany z miernictwem geologicznym. Dla każdego piksela obrazu mierzy się unikalne sygnatury spektralne obserwowanej powierzchni, wskazujące zależności współczynnika odbicia światła od długości fali [23]. Otrzymywane są w ten sposób trójwymiarowe struktury danych, które mogą być analizowane zarówno w dziedzinie spektralnej (λ) jak i przestrzennej (x, y). W zależności od liczby rozważanych kanałów oraz ciągłości odpowiadających im pasm wyróżnia się obrazowanie multispektralne (kilkanaście dyskretnych kanałów) oraz hiperspektralne (kilkadziesiąt i więcej kanałów, których pasma są na tyle blisko, że otrzymaną sygnaturę można traktować jako widmo ciągłe) [24]. Pomiar polowy może być efektem skanowania punkt po punkcie danej powierzchni badanej [25]. Procedura ta jest długotrwała i wymaga stabilności w czasie zarówno próbki, jak i źródła oświetlającego. Skanowanie próbki o nieregularnym kształcie jest utrudnione ze względu na konieczność ogniskowania się wiązki świetlnej w różnych odległościach od powierzchni światłowodu wprowadzającego sygnał do spektrometru. W przypadku tego typu próbek, jak również próbek o charakterystyce widmowej zmieniającej się w czasie pojedynczego pomiaru, lepiej sprawdza się rejestracja polowa zapewniająca jednoczesną akwizycję danych z całej powierzchni za pomocą macierzy detektorów (tzw. kamery multi- lub hiperspektralne [26]).

Innym rozwiązaniem jest wykorzystanie detektora typu kamera CCD rejestrującego pełne widmo w jednym wymiarze przestrzennym i przesuwnika taśmowego pozwalającego na rejestrację drugiego wymiaru przestrzennego obiektu [27].

W przypadku, gdy nie wymagana jest informacja spektralna zebrana z powierzchni obiektu, stosuje się pomiar punktowy. Wielu autorów twierdzi, że wystarczy zarejestrować spektrum w kilku punktach próbki, aby skutecznie przeprowadzić proces klasyfikacji [28, 29]. Liczne metody pomiarów spektralnych wykorzystujące inne zakresy spektralne niż VIS, umożliwiają akwizycje jedynie punktowo [30].

Opisane w niniejszym rozdziale techniki rejestracji widm mogą być zrealizowane w jednej z dwóch podstawowych konfiguracji układów optycznych: układ w świetle przechodzącym oraz układ w świetle odbitym. Cechą charakterystyczną pierwszego jest usytuowanie próbki na drodze optycznej pomiędzy źródłem światła a detektorem. W drugim przypadku rejestrator i źródło promieniowania znajdują się po tej samej stronie próbki. Rysunki poglądowe obu wymienionych konfiguracji przedstawione są na grafice 2.2. Pomiary związane z ekologią [31], geologią i dziedzinami pokrewnymi [32], podobnie jak ogólnie pojęte pomiary barwy bazują na rejestracji reflektancji. W dziedzinie dziedzictwa kulturowego można znaleźć prace opisujące wykorzystanie konfiguracji układu w świetle przechodzącym [33], jak i publikacje skupiające się na pomiarach w układzie światła odbitego [34, 35]. Typowym przykładem pomiarów transmitancji jest dziedzina chemii zajmująca się identyfikacją widm absorpcyjnych [36].

2.2. Przetwarzanie danych intensywnościowych oraz klasyfikacja

2.2.1. Ogólne podejście do procesu klasyfikacji

Podczas wykorzystania spektralnych technik pomiarowych, w tym technik VIS, sygnałem rejestrowanym jest intensywność. Jest ona przypisana konkretnym zakresom widmowym lub konkretnym długościom fal. Przeważnie techniki spektralne wiążą się z akwizycją dużej liczby danych. W przypadku pomiarów punktowych wielkość plików zależy przede wszystkim od rozdzielczości widmowej rejestrowanego sygnału, natomiast podczas rejestracji hiperspektralnych - wykorzystujących techniki powierzchniowe - ilość danych zależy od rozdzielczości widmowej i przestrzennej rejestrowanych obrazów. Podczas analizy tego typu danych, wskazane

jest zastosowanie cyfrowej analizy sygnału. Dopiero przetworzone dane mogą w skuteczny sposób zostać poddane dalszym procesom.

Klasyfikacją nazywa się podział obserwacji na klasy na podstawie cech tych obserwacji [37]. Proces przyporządkowania próbki do klasy, na podstawie pomiarów widmowych, można podzielić na trzy główne etapy: przetwarzanie wstępne, selekcja cech i klasyfikacja. Procedura pomiarowa rozpoczyna się zarejestrowaniem intensywności w dziedzinie widmowej oraz kończy przyporządkowaniem próbki do grupy (rysunek 2.1).

Podczas przetwarzania wstępnego można wyróżnić proces eliminacji wpływu charakterystyki podłoża i wykorzystywanego źródła światła, czyli inaczej mówiąc obliczenie transmitancji lub reflektancji badanego materiału. Mogą wystąpić dwie podstawowe sytuacje: próbka znajduje się na materiale zwanym bazą (jest to np. mikroskopowe szkiełko bazowe) lub próbka mierzona jest bezpośrednio.W momencie, gdy sygnał pochodzący od badanego materiału jest już pozbawiony wpływu elementów układu, następuje filtracja szumu. Często stosowanym filtrem w przypadku pomiarów spektralnych jest filtr Savitskiego-Golaya (filtr S-G).

Selekcja cech rejestrowanego sygnału jest wskazana do uzyskania skutecznej i nie nazbyt obciążającej obliczeniowo klasyfikacji badanych próbek. Jest ona jednoznaczna ze zmniejszeniem rozmiaru danych wprowadzanych następnie do klasyfikatora. Wykonuje się ją poprzez wygenerowanie nowego wektora cech. Obliczone cechy powstają na bazie danych wejściowych. Wygenerowany wektor cech może również być efektem redukcji niezmienionych danych lub łączyć dwa wymienione przypadki tworząc wektor zredukowanych nowych danych bazujących na pomiarach spektralnych. Przedstawicielami pierwszego sposobu są automatyczne techniki tj. liniowa analiza dyskryminacyjna (Linear Discriminant Analysis - LDA) [38], analiza składowych głównych (Principal Component Analysis - PCA) [39, 40], czy metoda cząstkowych najmniejszych kwadratów (Partial Least Square - PLS) [41]. Usuwanie danych odpowiadających konkretnym długościom fal lub całym zakresom spektralnym może być zaimplementowane w sposób manualny, lub automatyczny np. za pomocą algorytmów postępującej selekcji sekwencyjnej (Sequential Forward Selection - SFS), ruchomej selekcji postępującej (Sequential Floating Forward Selection - SFFS), czy wstecznej selekcji sekwencyjnej (Sequential Backward Selection - SBS) [42, 43].

Zarówno wygenerowanie nowego wektora cech, jak i usunięcie wyselekcjonowanych fragmentów widma, wykorzystane są w autorskich metodach parametryzacji i redukcji (rozdział 3).

W publikacjach z zakresu pomiarów spektralnych, autorzy opisują również inne techniki selekcji cech, takie jak: równoległa analiza czynnikowa (PARAllel FACtor analysis - PARAFAC) [44], współczynnik Fisher'a [45], czy analiza falkowa (Wavelet Transform - WT) [46].

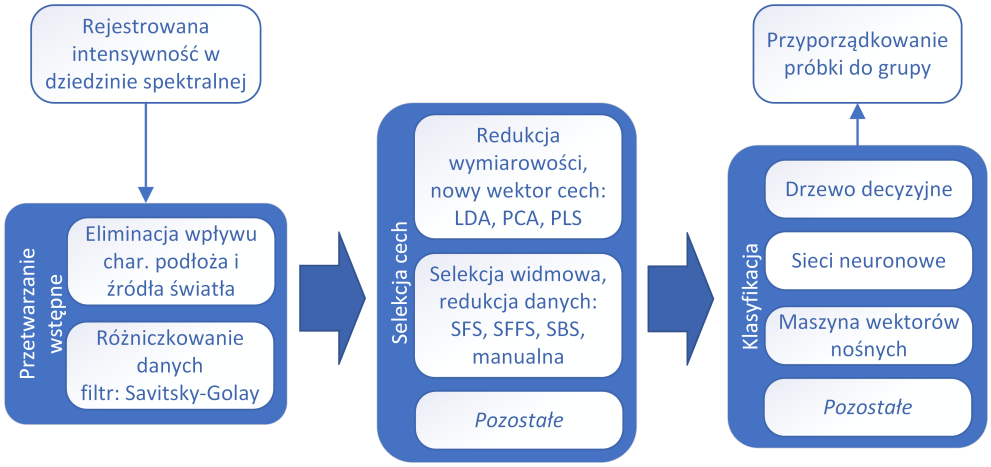
Etapem poprzedzającym końcowe przyporządkowanie próbki do odpowiedniej grupy jest klasyfikacja. Uczenie maszynowe, jak i metody statystyczne dają szerokie możliwości wyboru odpowiedniego klasyfikatora. Klasyczne podejście do klasyfikacji reprezentowane jest poprzez drzewo decyzyjne. Ta metoda, jako jedna z niewielu, umożliwia wgląd we wszystkie etapy klasyfikacji. Wiadomo, która cecha, na którym etapie, z jaką wagą przyczynia się do konkretnego podziału zbioru elementów. Takie podejście ułatwia interpretację wyników. Wykorzystanie sztucznych sieci neuronowych (Artificial Neural Network - ANN) [47], popularne w ostatnich latach, nie umożliwia tak łatwej interpretacji. Innym sposobem podejścia do zagadnienia klasyfikacji jest wykorzystanie maszyny wektorów nośnych (Support Vector Machine - SVM) [48].

Przykładami innych metod klasyfikacji są m. in.: analiza dyskryminacyjna cząstkowych najmniejszych kwadratów (Partial Least Square Prediction - Discriminant Analysis - PLS-DA) [44], hierarchiczna analiza klastrów (Hierarchical Cluster Analysis - HCA) [49], czy analiza wariancji kanonicznej (Canonical Variate Analysis - CVA) [50].

Autorzy wykorzystują do badań bardzo różne klasyfikatory. W wielu przypadkach ograniczają się do ich bezpośredniego zaimplementowania na danych pomiarowych, a uzyskane wyniki nie wskazują w sposób jawny, które zakresy widmowe niosły ze sobą cenną informację w konkretnym przypadku klasyfikacji. Oznacza to, że ewentualne późniejsze dostosowanie układu pomiarowego do konkretnego zastosowania nie jest w takim przypadku możliwe.

2.2.2. Wstępne przetwarzanie danych

We wstępnym przetwarzaniu danych rozróżnia się przypadek bezpośrednio mierzonej próbki i badanego materiału naniesionego na tzw. bazę, czyli np. na bazowe szkiełko mikroskopowe. Sposób umiejscowienia substancji badanej w układzie optycznym determinuje wykonanie odpowiedniej procedury eliminacji wpływu charakterystyki podłoża i źródła światła na rejestrowany sygnał. Tak obliczona transmitancja materiału poddawana jest procesowi filtracji szumów.



Rysunek 2.1: Schemat procesu klasyfikacji obiektów warstwowych wykorzystujący techniki spektralne VIS. „Char." – charakterystyki.

2.2.2.1. Eliminacja wpływu charakterystyki podłoża i źródła światła

Rejestrując wiązkę światła będącą efektem interakcji z badanym materiałem należy pamiętać, że otrzymywana informacja jest sumarycznym wynikiem oddziaływania wielu czynników. Właściwości optyczne elementów składowych układu takie jak charakterystyki spektralne, czy tłumienność, zawsze wywierają wpływ na wynik końcowy. Również kształt widma źródła światła oświetlającego próbkę determinuje wynikowy sygnał.

W układach transmisyjnych, najczęściej wykorzystywanymi wielkościami do opisu charakterystyk spektralnych badanych próbek są transmitancja i absorbancja w dziedzinie widmowej [10, 51–53]. Transmitancja definiowana jest, jako stosunek intensywności światła przechodzącego do intensywności światła padającego na próbkę (wzór 2.3) [54]. W analogiczny sposób, w układach odbiciowych, opisuje się reflektancję w dziedzinie widma wykorzystywaną do charakterystyki spektralnej obiektów słabo- lub nieprzepuszczalnych (wzór 2.4) [24, 55, 56].

W przypadku, gdy próbka nie wymaga nałożenia na podłoże, sygnał rejestrowany jest bezpośrednio (rysunek 2.2). W takiej sytuacji konieczne jest uniezależnienie sygnału jedynie od wykorzystywanego źródła światła. Oznacza to wyeliminowanie wpływu kształtu widma światła oświetlającego próbkę na rejestrowany sygnał badanego obiektu. Efekt ten jest uzyskiwany poprzez obliczenie ilorazu zarejestrowanego sygnału z umieszczoną próbką w torze optycznym I_p i bezpośredniego sygnału intensywnościowego użytego źródła światła I_{zr}. Opisana procedura wygląda tak samo w układzie światła przechodzącego (obliczana transmitancja próbki T_p) jak i w układzie światła odbitego (obliczana reflektancja próbki R_p) [57]. W celu pomiaru intensywności źródła światła w zależności od długości fali, należy umieścić rejestrator w torze

optycznym źródła w odległości umożliwiającej zarejestrowanie dobrej jakości danych. Szczegółowe informacje dotyczące umiejscowienia aparatu są zależne od konkretnego sprzętu.

$$T_p = \frac{I_p}{I_{zr}} \tag{2.3}$$

$$R_p = \frac{I_p}{I_{zr}} \tag{2.4}$$

Jeśli urządzenie ma możliwość rejestrowania szerszego zakresu widmowego niż to, w jakim pracuje źródło światła, informacje spektralne na temat próbki będą zawarte jedynie w zakresie pracy źródła światła.

W układzie światła przechodzącego rejestrowana wiązka jest przepuszczana przez próbkę. W przypadku próbki naniesionej na bazę (np. mikroskopowe szkiełko bazowe, punkt a) rysunku 2.3) rejestrowane widmo jest wynikiem oddziaływania zarówno bazy, jak i materiału badanego, na emitowaną wiązkę.

Transmitancja sumaryczna T_{sum} (materiału badanego i bazy) jest ilorazem zarejestrowanej intensywności sumarycznej I_{sum} i strumienia świetlnego docierającego do próbki – intensywności źródła I_{zr} (rysunek 2.3, wzór 2.5). Jednocześnie można przedstawić ją jako iloczyn transmitancji składowych: transmitancji bazy T_b i transmitancji próbki T_p (wzór 2.6).

$$T_{sum} = \frac{I_{sum}}{I_{zr}} \tag{2.5}$$

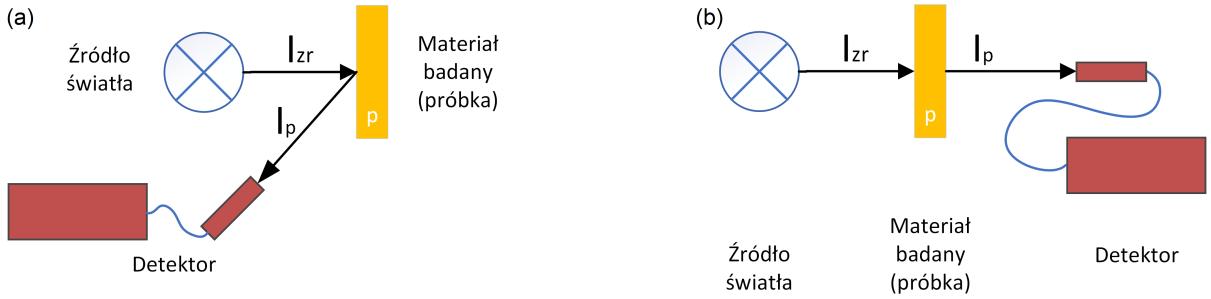
$$T_{sum} = T_b \cdot T_p \tag{2.6}$$

W celu uniezależnienia sygnału od bazy i widma użytego źródła światła należy zarejestrować intensywność wiązki po przejściu przez czystą bazę (bez nałożonego materiału do badań) I_b i powtórzyć procedurę obliczeniową, otrzymując transmitancję bazy T_b (wzór 2.7).

$$T_b = \frac{I_b}{I_{zr}} \tag{2.7}$$

Po wykonaniu przekształceń otrzymuje się transmitancję badanego materiału T_p (wzór 2.9).

$$\frac{I_{sum}}{I_{zr}} = \frac{I_b}{I_{zr}} \cdot T_p \tag{2.8}$$



Rysunek 2.2: Schematy układów do pomiaru (a) reflektancji, (b) transmitancji próbki, bez wykorzystania bazy. I_{zr} – intensywność światła docierającego do próbki, I_p – intensywność zarejestrowana po odbiciu lub przepuszczeniu przez materiał badany.

$$T_p = \frac{I_{sum}}{I_b} \tag{2.9}$$

W przypadku rejestrowania widma, powyższe przekształcenia należy rozpatrywać dla pojedynczych zakresów spektralnych (wzór 2.10).

$$T_p(\lambda) = \frac{I_{sum}(\lambda)}{I_b(\lambda)} \tag{2.10}$$

Gdzie λ oznacza długość fali lub zakres spektralny któremu odpowiada dana zmierzona wartość intensywności. Również odnosi się to do reflektancji (wzór 2.11).

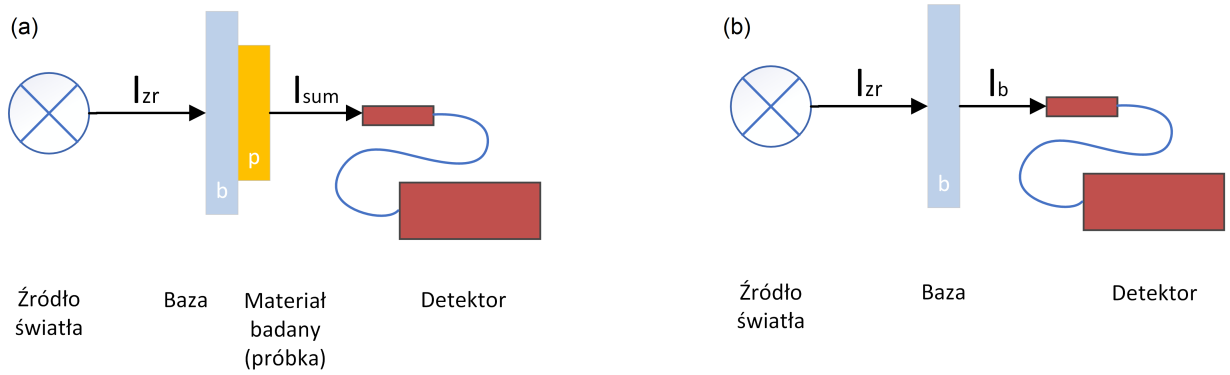
$$R_p(\lambda) = \frac{I_p(\lambda)}{I_{zr}(\lambda)} \tag{2.11}$$

W przypadku pomiaru substancji będącej wewnątrz innego materiału, referencją przez którą należy podzielić zarejestrowany sygnał jest intensywność po przejściu przez element niezawierający w danym momencie materiału badanego. Przykładem opisanej sytuacji jest badanie zawartości jaja, gdzie referencją będzie pojedyncza skorupa (wydmuszka) [58].

Wielokrotnie charakterystyki widmowe materiałów przedstawiane są w formie absorbancji A (inaczej tłumienie). Definiuje się, jako ujemną wartość logarytmu dziesiętnego z transmitancji [54] lub reflektancji [59] (wzór 2.12).

$$A = -log_{10}T = -log_{10}R \tag{2.12}$$

Transmitancja zmienia się liniowo (0–100%), natomiast absorbancja logarytmicznie (0–∞). W przypadku idealnym, zachodzi zależność opisywana przez prawo Lamberta–Beera. Stanowi ono o tym, iż tłumienie wiązki świetlnej w materiale jednorodnym zmienia się wykładniczo



Rysunek 2.3: Schemat układu do pomiaru (a) transmitancji próbki wraz z bazą, (b) intensywności wiązki po przejściu przez bazę – bez nałożonego materiału testowanego. I_{zr} – intensywność światła docierającego do bazy, I_{sum} – sumaryczna intensywność zarejestrowana po przejściu przez bazę i materiał badany, I_b – intensywność zarejestrowana po przejściu przez samą bazę.

i rośnie wraz z drogą przejścia promieni przez substancję. Jednocześnie prawo to opisuje fakt mówiący o tym, że każda warstwa tej samej grubości absorbuje taką samą część energii niezależną od intensywności padającego promieniowania pod warunkiem, że promieniowanie nie zmienia stanu fizycznego lub chemicznego ośrodka, co dowodzi o liniowości transmitancji [60]. W przypadku zastosowania logarytmu naturalnego, mówi się o ekstynkcji E promieniowania elektromagnetycznego (wzór 2.13) będącej alternatywą prezentacji charakterystyki spektralnej badanego materiału [61].

$$E = -lnT \tag{2.13}$$

W przedstawionych rozważaniach nie zostały wzięte pod uwagę aspekty ilościowe promieniowania uzależnione od kątów padania i rejestracji strumieni świetlnych, jak również od struktury materiałów mierzonych świadczących o charakterze dyfuzyjnym lub kierunkowym przepuszczania lub odbicia promieni, co szczegółowo opisane jest np. w pozycji [62]. Podczas analizy spektralnej próbek można ograniczyć się do rejestracji widma nie uwzględniając relacji intensywnościowych pomiędzy zbadanymi próbkami. To podejście zostało wykorzystane w rozprawie.

2.2.2.2. Operacja filtracji

Odpowiednie przygotowanie sygnału do dalszej obróbki i późniejsze zaimplementowanie go do wybranego klasyfikatora, nie jest oczywiste. Nieprawidłowe dobranie sposobu filtracji

lub niedopasowanie parametrów konkretnej metody obróbki, może spowodować usunięcie istotnych danych lub pozostawić szumy.

Podczas przetwarzania wstępnego wyróżnia się proces uniezależnienia sygnału od podłoża i źródła światła oraz korekcję rozproszenia i różniczkowanie widma. Multiplikatywna korekcja rozproszenia (Multiplicative Scatter Correction MSC) [63] i jej modyfikacje (przedstawione m.in. w pozycji [64]) lub transformacja SNV (Standard Normal Variate) [65] są przykładami algorytmów pozwalających wyeliminować niekorzystne efekty fizyczne zachodzące podczas rejestracji pomiarów w układach światła odbitego – takie jak rozproszenie. W podstawowej formie MSC polega na rejestracji widma referencyjnego (X_r), niezawierającego składowej rozproszonej, a następnie obliczenie regresji liniowej każdego pomiaru (X_i) w stosunku do referencyjnego (wzór 2.14). Spektrum po korekcie (X_{cor}) obliczane jest na podstawie wzoru 2.15. Główna różnica między MSC, a SNV polega na tym, że transformacja SNV nie wymaga pomiaru referencyjnego, a korekcja jest wykonywana na każdym spektrum osobno.

$$X_i \approx a_i + b_i X_r \tag{2.14}$$

$$X_{cor} = \frac{X_i - a_i}{b_i} \tag{2.15}$$

Różniczkowanie sygnału jest znaną od dawna metodą na usunięcie przesunięcia addytywnego (tzw. linii bazy, przesunięcia funkcji o stałą), jak również efektu multiplikacyjnego zarejestrowanej absorbancji [66]. Realizuje się to poprzez wykorzystanie różniczkowania Norrisa–Williamsa N-W [67] oraz różniczkowo wielomianowy filtr Savitskiego–Golaya S-G [68]. Pierwsza pochodna pozwala określić szybkość zmian, ponadto jej wartość osiąga zero dla długości fali odpowiadającej przypadającemu maksimum absorbancji. Ekstrema pierwszej pochodnej odpowiadają punktom przegięcia zarejestrowanej funkcji. Drugą pochodną charakteryzuje minimum lokalne, którego odpowiadająca długość fali przypada na maksimum absorbancji. W kolejnych parzystych stopniach pochodnej następuje podobny efekt – ekstremum pochodnej odpowiada długości fali, dla której występuje maksymalna absorpcja [69]. Dziedzina, w której analizuje się pochodne zarejestrowanych absorbancji nazywana jest spektroskopią różnicową. Wykorzystuje się ją przede wszystkim w chemii analitycznej i naukach pokrewnych. Sygnały pochodzą z prześwietlanych substancji będących przeważnie w stanie ciekłym, umieszczonych w urządzeniach zwanych spektrofotometrami. Warunki, które

są zapewniane podczas badania, umożliwiają uzyskanie sygnału o bardzo niskim poziomie szumów. Wykonanie różniczkowania widma, zarejestrowanych tym sposobem absorbancji, powoduje wzrost selektywności i czułości pomiaru spektrofotometrycznego w porównaniu do klasycznego podejścia. Zatem możliwa jest wieloskładnikowa analiza mieszanin czynników o zbliżonych do siebie widmach.

Najprostszym podejściem do obliczenia pochodnej jest wykorzystanie metody różnic skończonych – pierwsza pochodna obliczana jest, jako różnica intensywności między dwoma kolejnymi spektralnymi punktami pomiarowymi. Druga pochodna analogicznie – w obliczeniach używa się danych uzyskanych po pierwszym różniczkowaniu. Takie podejście można stosować na teoretycznych przebiegach, ewentualnie po wcześniejszym wygładzeniu widma [70, 71], jednak w przeważającej ilości analiz spektralnych, obliczanie różnic skończonych zbyt mocno wzmacnia niepożądane, wysokie częstotliwości i z tego powodu w praktyce jest rzadko wykorzystywane.

Użycie filtru N-W lub S-G skutecznie przeciwdziała znacznemu pogorszeniu się stosunku sygnału do szumu (signal-to-noise ratio SNR) podczas różniczkowania sygnału. W obu przypadkach najpierw sygnał jest wygładzony, następnie obliczana jest pochodna. W ogólności, zasadę działania filtra można zapisać za pomocą wzoru 2.16, gdzie $f_i \equiv f(t_i)$ są równomiernie rozmieszczonymi danymi w czasie $t_i \equiv t_0 + i\triangle$, dla stałych odstępów $\triangle$ oraz $i = ..., -2, -1, 0, 1, 2, ...$ Szerokość okna filtracji opisywana jest liczbą punktów „z lewej strony" – n_L oraz liczbą punktów „z prawej strony" – n_R. Zadaniem filtra jest znalezienie takich współczynniki c_n, które maksymalizują dopasowanie efektu działania funkcji do zadanych danych.

W literaturze dotyczącej pomiarów spektralnych częściej występuje opis zastosowania filtru S-G niż N-W. W przypadku filtra S-G dla każdego punktu f_i, do wszystkich $n_L + n_R + 1$ punktów ruchomego okna dopasowywany jest wielomian za pomocą metody najmniejszych kwadratów. Użytkownik decyduje o stopniu wielomianu oraz o szerokości okna. Następnie wielkość g_i zostaje zapisana, jako wartość tego wielomianu w pozycji i. Po przesunięciu okna do następnej pozycji f_{i+1} procedura dopasowania wykonywana jest od początku. Stosowane metody numeryczne pozwoliły na uproszczenie koniecznych obliczeń i opracowanie zestawów współczynników c_N dla których równanie 2.16 „automatycznie" realizuje proces dopasowania

wielomianów za pomocą algorytmu najmniejszych kwadratów do ruchomego okna [72].

$$g_i = \sum_{n=-n_L}^{n_R} c_n f_{i+n} \tag{2.16}$$

Filtr S-G powszechnie stosuje się jako filtr wygładzający do sygnałów wolnozmiennych, zaszumionych wysokimi częstotliwościami. Zastosowanie tego filtru na danych spektralnych daje lepsze rezultaty, niż typowy filtr uśredniający o zmiennym oknie [72]. Filtry N-W i S-G, nie eliminują z sygnału składowej pochodzącej od rozproszenia światła na próbce, jednak można uznać, że po ich zastosowaniu jej wpływ zostaje zmniejszony [69].

2.2.3. Selekcja cech

W przypadku, gdy obiekt opisywany jest przez dziesiątki, setki a nawet tysiące atrybutów, mówi się o danych wielowymiarowych. Takie zbiory generują problemy podczas analizy. W przestrzeniach wielowymiarowych obiekty mogą okazać się ciężej klasyfikowalne. Występuje zjawisko nazwane "przekleństwem wymiarowości" [73] głoszące, iż liczba próbek musi rosnąć wykładniczo wraz ze wzrostem wymiaru przestrzeni cech. Wymagane jest znaczne zwiększenie liczby obiektów mierzonych, co w wielu badaniach jest niemożliwe. Oznacza to, że redukcja cech czasem może okazać się konieczna do przeprowadzenia poprawnej klasyfikacji [74, 75].

2.2.3.1. Redukcja wymiarowości - generacja nowych cech

Poprzez użycie sformułowania redukcja wymiarowości przestrzeni poprzez ekstrakcję cech, należy rozumieć taką transformację wektora wejściowego cech, która generuje zupełnie nowy, zmniejszony wektor atrybutów niepokrywających się z początkowymi. Nowe cechy mogą być wyselekcjonowane na podstawie kryterium reprezentatywności (najlepiej odwzorowujący opis danych wejściowych w nowej przestrzeni) lub separowalności klas. Zadanie generacji nowych cech polega na wykryciu wewnętrznej struktury zbioru danych lub współzależności miedzy tymi danymi [76].

Najstarszą, ale ciągle wykorzystywaną przez naukowców [38, 77, 78], metodą automatycznej redukcji wymiarów jest liniowa analiza dyskryminacyjna (Linear Discriminant Analysis – LDA), również często uznawana za klasyfikator sam w sobie. Jej autorem jest R. Fisher, który w 1936 r. zaproponował podejście mające na celu znalezienie liniowej kombinacji cech najlepiej rozdzielającej dwie klasy. Wynik poszukiwania zależy od wyboru wartości

początkowej wektora wag, dlatego procedurę optymalizacji można przeprowadzić kilka razy, dla różnych, losowo wybranych wektorów wag początkowych, a następnie wybrać najlepsze rozwiązanie [79]. LDA można uogólnić do większej liczby klas, co opisują w swoich pracach C. R. Rao [80] i J. G. Bryan [81], jak również m. in. W. Malina [82], czy A. Kołakowska [83].

Najpopularniejszym algorytmem do automatycznej redukcji wymiarów jest analiza składowych głównych (Principal Component Analysis – PCA) [84, 85]. Większość autorów przygotowuje sygnał do procesu redukcji wymiarowości za pomocą filtracji, usuwania składowej stałej, czy standaryzacji danych [86–88], jednakże spotyka się publikacje, gdzie PCA wykonywane jest na surowych danych [29, 59]. PCA pozwala wykonać rzut wielowymiarowych danych na przestrzeń o dużo mniejszym wymiarze, jednocześnie zachowując maksymalnie dużo informacji [89].

Nierzadko autorzy, w celu zmniejszenia wymiarowości danych, wykorzystują metodę cząstkowych najmniejszych kwadratów (Partial Least Squares – PLS) [90]. Ma ona za zadanie znaleźć liniową zależność pomiędzy macierzą próbek, a wektorem przynależności do klas oraz wskazać tzw. komponenty (zmienne) ukryte, których liczbę wskazuje operator. W tym celu maksymalizuje funkcję celu, twierdzącą o maksymalizacji kowariancji pomiędzy liniową kombinacją wektorów macierzy próbek, a macierzą odpowiedzi. Jej pierwotną wersją był algorytm NIPALS (Nonlinear Iterative Partial Least Squares) [91]. Ideą współczesnego NIPALS jest dekompozycja biliniowa [92]. To dzięki niej, otrzymywana jest macierz komponentów przekształcająca macierz próbek na macierz odpowiedzi. Powstało wiele modyfikacji algorytmu NIPALS, w efekcie również PLS.

Metodę PLS stosuje się przede wszystkim na danych wysoko skorelowanych, jak również wtedy, gdy liczba zmiennych znacznie przewyższa liczbę próbek [93, 94]. PLS krytykowana jest za trudny do interpretacji model. Wyniki w postaci zależności pomiędzy obliczonymi na jej podstawie predyktorami oraz macierzą odpowiedzi również są niełatwe do interpretacji. Ustalenie większej liczby komponentów ukrytych powoduje wzrost złożoności modelu, jak również ma wpływ na jego sprawność.

Podsumowanie właściwości podstawowych metod redukcji wymiarowości generujących macierze nowych cech przedstawiono w tabeli 2.1.

Tabela 2.1: Podsumowanie i porównanie cech podstawowych metod redukcji wymiarowości generujących macierze nowych cech. LDA – liniowa analiza dyskryminacyjna, PCA – analiza składowych głównych, PLS – metoda cząstkowych najmniejszych kwadratów.

	możliwość wykonania gdy liczba atrybutów > liczba próbek	możliwość kontroli procesu	interpretowalność wyselekcjonowanych cech	dane nie muszą być ze sobą skorelowane
LDA	✗	✗	✗	✓
PCA	✗	✗	✗	✗
PLS	✓	w małym zakresie	✗	zazwyczaj ✗

Ze względu na powszechność wykorzystania, wybrano metodę PCA jako referencję dla rozwiązań zaproponowanych w rozprawie. Szczegółowy opis działania PCA przedstawiono poniżej.

Nowy wektor, powstały po zastosowaniu tej metody, składa się z tzw. składowych głównych – funkcji liniowych zmiennych oryginalnych. Ich zadaniem jest wyjaśnienie w maksymalnym stopniu całkowitej wariancji danych wejściowych. Pierwsza składowa główna opisuje w największym stopniu zróżnicowanie danych i może być ona rozpatrywana jako prosta, na którą rzutowane są ortogonalnie wartości danych wejściowych. Prosta ta jest dopasowana w ten sposób, aby suma kwadratów odległości punktów danych początkowych była od niej minimalna. Kolejne składowe opisują wariancję odpowiednio w coraz mniejszym stopniu. Pierwsza i druga składowa główna tworzą ”najlepiej” dopasowaną płaszczyznę w sensie: suma kwadratów odległości prostopadłych punktów danych początkowych od tej płaszczyzny, jest minimalna [76]. Korzystanie jedynie z kilku pierwszych wektorów głównych może doprowadzić do zmniejszenia poziomu szumu danych, które są zazwyczaj opisywane przez dalsze składowe.

Przy założeniu, że dane wejściowe są w postaci macierzy $X[m, n]$, gdzie m to liczba cech, a n to liczba próbek procedurę obliczeń PCA można przedstawić jako:

1. obliczenie wektora średnich dla wierszy macierzy X, czyli średnie wartości kolejnych cech dla wszystkich obserwacji;

$$u[m] = \frac{1}{N} \sum_{n=1}^{N} X[m, n] \tag{2.17}$$

2. obliczenie odchyleń od średniej;

$$B[m,n] = X[m,n] - u[m] \tag{2.18}$$

3. obliczenie macierzy kowariancji na podstawie B

$$C[n,n] = \frac{1}{n-1} B^T B \tag{2.19}$$

   postać dzielnika jest w formie $n-1$, a nie n ze względu na poprawkę Bessala;
4. obliczenie macierzy wektorów osobliwych V na podstawie macierzy C

$$V^{-1} C V = D \tag{2.20}$$

   gdzie: D jest macierzą diagonalną [n,n] składowych głównych C; V jest macierzą [n,n] o n wektorach osobliwych (kolumny V) macierzy C;
5. wybór maksymalnych wartości składowych głównych z macierzy D – minimalizacja strat podczas rzutowania na przestrzeń o mniejszej liczbie wymiarów;
6. wybór L wektorów osobliwych z macierzy V odpowiadających wybranym wartościom macierzy D;

$$V[n,n] \Rightarrow W[n,L] \tag{2.21}$$

7. rzutowanie wektorów osobliwych na przestrzeń o mniejszej liczbie wymiarów

$$Y = V^T x \tag{2.22}$$

   gdzie: V jest macierzą wektorów osobliwych; x jest rzutowanym wektorem osobliwym odpowiadającym kolumnie macierzy W.

Jak każda metoda, PCA ma również swoje ograniczenia. W przypadku, gdy jest mniej próbek niż atrybutów, wykonanie tej metody jest technicznie niemożliwe. Stosując pewne założenia możliwe jest uzyskanie wyników, ale nie należy być przekonanym o ich słuszności. Ponadto, algorytm transformuje dane z dziedziny widmowej, gdzie możliwa jest interpretacja fizyczna, do nowej przestrzeni obserwacji, która już tego nie zapewnia. Jest to silny argument stanowiący o tym, że PCA nie jest idealną metodą podczas badań próbek fizycznych. Co więcej, metoda nie umożliwia użytkownikowi kontroli procesu obliczeniowego. Proces może

być traktowany jako "czarna skrzynka". Ta cecha może być również traktowana jako zaleta – nie wymaga doświadczenia operatora, jednak od strony naukowej, może być postrzegane jako pewien mankament. W przypadku danych nieskorelowanych nie należy wykonywać PCA [76].

2.2.3.2. Redukcja wymiarowości - selekcja widmowa

Selekcja widmowa polega na zmniejszeniu liczby analizowanych kanałów spektralnych poprzez stworzenie takiego wektora cech, który będzie zawierał możliwie mały zbiór cech przy jednoczesnej optymalizacji zadanej funkcji kryterialnej. Atrybuty są identyczne, jak w zbiorze początkowym, jedynie zmniejszona zostaje ich liczba. Podczas procesu doboru cech sprawdzany jest wynik klasyfikacji za pomocą wcześniej wskazanego klasyfikatora.

Do tego zagadnienia można podejść w sposób niecyfrowy – manualna selekcja kanałów [95] – stosowany przede wszystkim w pomiarach hiperspektralnych np. w kartografii. Możliwa również jest automatyczna selekcja kanałów za pomocą takich algorytmów jak: przeszukiwanie wyczerpujące [96] – algorytm wykładniczy, selekcja postępująca/wsteczna (Sequential Forward/Backward Selection SFS/SBS) – algorytm sekwencyjny [97, 98], czy algorytm stochastyczny [99].

Zasada działania algorytmu sekwencyjnego polega na dodawaniu, lub usuwaniu cech ze zbioru tak długo, aż funkcja celu osiągnie maximum. W przypadku selekcji postępującej proces rozpoczyna się od zbioru pustego, następnie wprowadzane są do niego najistotniejsze cechy i sprawdzana jest wartość sprawności klasyfikatora. W momencie, gdy dodana cecha spowoduje największy przyrost wartości funkcji kryterialnej lub zbiór osiągnie zamierzoną liczbę cech, następuje zakończenie procesu. Algorytm selekcji wstecznej działa w sposób analogiczny, rozpoczynając proces od zbioru wszystkich cech i usuwając kolejno atrybuty o najmniejszym znaczeniu. Automatyczne algorytmy sekwencyjne charakteryzują się prostotą działania, jednak istnieją przypadki, w których będą cechować się wysoką złożonością. Główną ich wadą jest brak możliwości usunięcia raz dodanej cechy ze zbioru. Może nastąpić sytuacja, w której jedna z wcześniej dodanych cech, spowoduje zablokowanie przyrostu wartości funkcji kryterialnej. W takim przypadku, dodanie kolejnego atrybutu mającego potencjał polepszenia wyniku, nie przyniosłoby zamierzonego skutku. Wada ta została usunięta w algorytmie ruchomej selekcji postępującej (Sequential Floating Forward Selection – SFFS), która po każdorazowej iteracji algorytmu wykonuje serię warunkowych usunięć cech z opracowywanego zbioru. Prowadzi to do skuteczniejszej selekcji widmowej, jednak złożoność obliczeniowa algorytmu znacznie przewyższa złożoność pozostałych.

Należy zwrócić uwagę, że opisywane automatyczne algorytmy świetnie sprawdzają się w dziedzinach bazujących przede wszystkim na danych numerycznych, takich, jak matematyka czy technologia informacyjna. W zastosowaniach inżynierskich, czy biologicznych, gdzie wynikiem działania algorytmu mogą być pojedyncze długości fal nie współtworzące konkretnego kanału spektralnego, może okazać się, że ta metoda nie jest optymalna. Selekcja pojedynczych cech, oznaczająca wyodrębnienie pojedynczych długości fal, lub ich wąskich zakresów, w przypadku pomiarów spektralnych, z dużym prawdopodobieństwem nie będzie dawać możliwości interpretacji fizycznej. Przekłada się to na brak możliwości dopasowania układu do konkretnego zastosowania.

Podsumowanie właściwości podstawowych metod selekcji widmowej przedstawiono w tabeli 2.2. Prezentowane metody wymagają zaimplementowania również klasyfikatora. W podsumowaniu wykorzystano nieskomplikowany klasyfikator – drzewo decyzyjne (decision tree – DT), którego zasada działania przedstawiona jest w rozdziale 2.2.4.

Tabela 2.2: Podsumowanie i porównanie cech podstawowych metod automatycznej selekcji widmowej (SFS/SBS – selekcja postępująca/wsteczna i SFFS – ruchoma selekcja postępująca) oraz nieautomatycznej (selekcja ręczna i autorska metoda redukcji). Porównanie cech algorytmów zaprezentowano na przykładzie połączenia ich z klasyfikatorem DT – drzewo decyzyjne.

	złożoność obliczeniowa 1–3 (mała–duża)	czas wykonania procesu 1–3 (krótko–długo)	interpretowalność wyselekcjonowanych cech	niewymagane doświadczenie operatora	możliwość kontroli procesu
SFS/SBS + DT	2	1	✗	✓	✗
SFFS + DT	3	2	✗	✓	✗
selekcja manualna + DT	1	3	✓	✗	✓

2.2.4. Klasyfikacja w uczeniu statystycznym

Uczenie statystyczne, łączy w sobie elementy uczenia maszynowego oraz, dużo starszej i dłużej rozwijającej się dziedziny – statystyki. Uczenie maszynowe skupia się przede wszystkim na zdolności predykcyjnej zaproponowanego modelu. Natomiast w statystyce najważniejszym jest wyjaśnienie i zrozumienie procesu, który doprowadził do powstania określonych zależności w zbiorze danych. W podejściu statystycznym dużą rolę przykłada się do odpowiedniego zebrania próby, poprawności losowania, reprezentatywności, co nie jest brane pod uwagę w uczeniu maszynowym, gdzie operuje się na danych zastanych. Ze względu

na komplementarne cechy uczenia maszynowego i statystyki, połączenie ich tworzące uczenie statystyczne daje dużo szersze możliwości.

Jednym z podstawowych i jednocześnie bardzo obszernych działów uczenia statystycznego, jest klasyfikacja obiektów [100, 101]. Aby w skuteczny sposób wykorzystać algorytmy klasyfikacyjne, należy zdefiniować wartości zmiennej objaśnianej oraz zmienną kontrolną. Model uczenia statystycznego ma za zadanie estymować wartości zmiennej objaśnianej. Mogą być one zarówno ciągłe np. czas nałożenia substancji na bazę, jak i należące do zbioru dyskretnego np. klasyfikacja do grup „zdrowy", „chory". Zmienna kontrolna jest stała podczas pomiaru, wprowadza się ją do modelu w celu ustalenia, czy nie wpływa na zależność pomiędzy innymi analizowanymi zmiennymi.

Dziedzinę klasyfikacji można podzielić na dwie główne grupy: pod nadzorem oraz nienadzorowaną. Uczenie pod nadzorem (klasyfikacja nadzorowana), to inaczej uczenie się z przykładów, na bazie dostępnych danych wejściowych, oraz wyjściowych. Proces uczenia jest możliwy dzięki klasyfikatorowi, będącemu pewną regułą klasyfikacyjną służącą do predykcji klasy, do której należy obserwacja [76]. Klasyfikatory dzieli się na parametryczne i nieparametryczne. Te pierwsze bazują na statystycznym prawdopodobieństwie rozkładu wzorców dla danej klasy. Przykładem tego typu klasyfikatora są drzewa decyzyjne. Klasyfikatory nieparametryczne nie wymagają założeń odnośnie rozkładu populacji, z której losowana jest próba, wykorzystują inne metody podziału klas jak np. regresja, czy sztuczne sieci neuronowe [102]. Podczas uczenia się bez nadzoru, danymi są jedynie dane wejściowe, niedostępny jest zbiór danych wyjściowych (brak wiedzy nt. struktury klasowej w tym zbiorze). W takim przypadku klasyfikacja polega na grupowaniu obiektów na podstawie wykrycia wewnętrznej struktury zbioru danych lub współzależności między nimi [103].

Niniejszy podrozdział poświęcony jest klasyfikacji uczenia pod nadzorem. Przedstawiono w nim kilka podstawowych klasyfikatorów – funkcji odwzorowujących przestrzeń cech w zbiór numerów klas. Algorytmy klasyfikacyjne wykorzystują podczas nauki jedynie dane z tzw. zbioru uczącego, który jest podzbiorem zbioru obiektów. Elementy będące w zbiorze uczącym składają się z wektora cech opisującego dany obiekt oraz z etykiety klasy, do której przynależą. Zadaniem klasyfikacji jest przydzielenie wektora cech do klasy, dla której dobierana funkcja klasyfikacyjna osiągnie swoje maksimum. Wynikiem działania algorytmu będzie indeks funkcji, dla której osiągnięto największą wartość. Proces klasyfikacji powinien zostać zakończony testem, podczas którego próbki nie będące w zbiorze uczącym podlegają

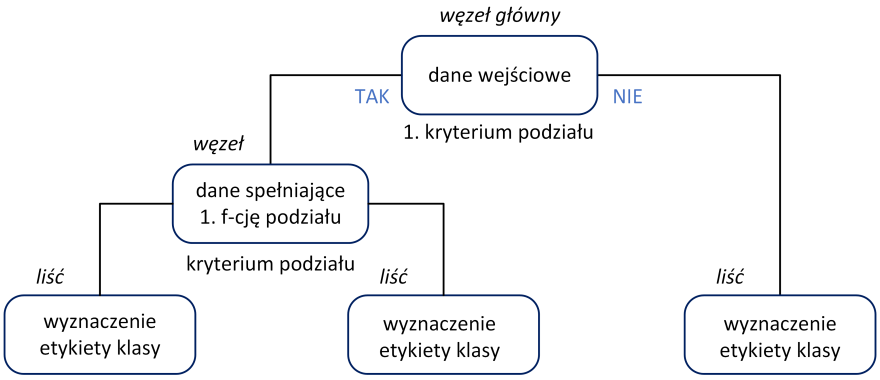
klasyfikacji z wykorzystaniem stworzonej wcześniej funkcji. Klasyfikacje nadzorowane są wrażliwe na strukturę danych użytych do ich uczenia [104], oznacza to, że skuteczność klasyfikatora będzie się zmieniała wraz z przestawieniem danych zarówno w zbiorze uczącym jak i testowym.

Klasyfikację pod nadzorem można rozpatrywać na różne sposoby: w kategoriach probabilistycznych (klasyfikator Bayesowski, Gaussowski, funkcje dyskryminacyjne i in.), metodą najbliższego sąsiada, wygenerowaniem drzewa decyzyjnego, wykorzystaniem sieci neuronowych, czy użyciem maszyny wektorów nośnych.

Klasyfikatory statystyczne i minimalnoodległościowe

Dla klasyfikatorów statystycznych przestrzenią obserwacji jest zbiór wszystkich możliwych wartości wektora cech, natomiast przestrzenią decyzyjną (iloczyn kartezjański obiektów istotnych ze względu na proces decyzyjny) zbiór wszystkich klas. Zakłada się wzajemną niezależność zmiennych niezależnych. Przykładem tej grupy jest optymalny klasyfikator Bayesa przyporządkowujący obiekt do klasy, dla której wartość prawdopodobieństwa a posteriori jest największa. Wymaga on znajomości rozkładu zmiennej losowej (prawdopodobieństwa pojawiania się obiektów z poszczególnych klas) i warunkowej gęstości rozkładu prawdopodobieństwa cech w klasach [105]. Założenia te trudno spełnić w praktycznych sytuacjach. Modyfikacje tego klasyfikatora rozwiązują zadany problem. Przykładem może być algorytm wykorzystujący PCA oraz klasyfikator Bayesa stworzony do klasyfikacji produktów ropy naftowej za pomocą spektroskopii NIR działający w czasie rzeczywistym [106].

Głównym przedstawicielem klasyfikatorów minimalnoodległościowych jest metoda najbliższego sąsiada (k-NN). Reprezentuje ona jedną z najważniejszych nieparametrycznych metod klasyfikacji. Obiekt przypisywany jest do tej klasy, do której należy większość z jego k sąsiadów. Metoda ma bardzo wysoką efektywność w przypadku wzrostu liczby obserwacji do nieskończoności [107], jednak przy ograniczonej liczbie próbek, jej efektywność drastycznie spada. Metodę k-NN z powodzeniem wykorzystuje się w spektroskopii MR w analizie obrazów, gdzie algorytm uczy się rozpoznawać różne tkanki poprzez początkowe wskazanie typów. Możliwa jest również taka modyfikacja algorytmu, która pozwala na wykonanie przyporządkowania w pełni automatycznie, z porównywalną dokładnością do technik manualnych [108]. K-NN może być również wykorzystywana w połączeniu z innymi metodami, jak np. maszyna wektorów nośnych w zastosowaniu klasyfikacji obrazów hiperspektralnych [109].



Rysunek 2.4: Graficzna wizualizacja idei binarnego drzewa decyzyjnego.

Drzewo decyzyjne

Ogólna idea budowy drzew klasyfikacyjnych polega na sekwencyjnym dzieleniu podzbiorów przestrzeni próby (węzłów) danych wejściowych na dwa (w przypadku metody CHAID – Chi-squared Automatic Interaction Detector – podział może być liczniejszy) rozłączne i dopełniające się podzbiory, rozpoczynając od całego zbioru danych wejściowych. Każdy węzeł (oprócz tych odpowiadających podzbiorom końcowym, zwanych liśćmi) zawiera funkcję określającą warunek podziału. W przypadku drzew binarnych funkcja ta przyjmuje jedną z dwóch wartości: prawda lub fałsz. Celem drzewa klasyfikacyjnego jest wyznaczenie funkcji przyporządkowującej każdemu elementowi zbioru liści dokładnie jedną etykietę klasy. Schemat konstrukcji drzewa decyzyjnego na przykładzie drzewa binarnego przedstawiony jest na rysunku 2.4.

Podczas konstrukcji drzewa decyzyjnego wyodrębnia się trzy podstawowe etapy:

1. dobór optymalnej metody podziału węzłów;
2. określenie wielkości drzewa, czyli określenie, kiedy algorytm ma zakończyć działanie;
3. wyznaczenie sposobu przyporządkowania etykiety klasy.

W pierwszym etapie algorytm rozdziela obserwacje zbioru uczącego należące do danego węzła na dwa podzbiory. Są one możliwie jednorodne ze względu na etykiety klas. Określana jest tzw. miara niejednorodności elementów w tym węźle. Może ona bazować na błędzie klasyfikacji, na funkcji entropii lub na funkcji zwanej indeksem Giniego [110]. Dobranie optymalnego podziału węzłów często wymaga rozważenia dużej ilości różnych podziałów ($2^{L-1} - 1$, gdzie L liczba różnych wartości cechy jakościowej). Wykorzystanie kryterium entropii lub indeksu Giniego, w zagadnieniach dwuklasowych ogranicza liczbę przeszukiwanych

podziałów do $L-1$. Wybór optymalnego podziału węzła, jest równoważny wyborowi podziału minimalizującego miarę niejednorodności drzewa klasyfikacyjnego [76].

Określenie wielkości drzewa wiąże się z podaniem reguły twierdzącej o zaprzestaniu „rozrastaniu" się drzewa po osiągnięciu konkretnego kryterium. W ten sposób ustala się, czy dany węzeł ma zostać liściem drzewa. Niewskazanie tego kryterium prowadzi do przetrenowania modelu. W takim przypadku obiekty z próby uczącej klasyfikowane są z minimalnym błędem, natomiast dane testowe cechują się dużo niższą dokładnością poprawnych klasyfikacji. Optymalna wielkość drzewa może być wyznaczona za pomocą reguły stopu (po określonej liczbie podziałów, dany węzeł uznawany jest za końcowy), lub po uzyskaniu odpowiedniej jednorodności drzewa, lub po wstępnym wygenerowaniu drzewa maksymalnego, a następnie jego selektywnym przycinaniu. Przycinanie drzewa klasyfikacyjnego ma za zadanie zminimalizować prawdopodobieństwo błędnego przyporządkowania i polega na wyeliminowaniu podziałów, które nie mają istotnego znaczenia dla poprawności klasyfikacji.

Wyznaczenie etykiety klasy jest efektem działania funkcji określonej na liściach drzewa. Różne modele drzew klasyfikacyjnych reprezentują różne tego typu funkcje.

Jednym z najstarszych modeli drzew klasyfikacyjnych jest model CHAID zaproponowany przez Kassa w 1980 r. [111, 112]. Za pomocą nieskomplikowanego algorytmu budowane jest drzewo, z którego węzłów mogą wychodzić więcej niż dwie gałęzie. Podstawą jego działania jest test Chi-kwadrat lub test F. Przede wszystkim przeznaczony jest do analizy dużych zbiorów danych tj. badań marketingowych, czy segmentacji rynku [113]. Jednak zdarzają się również zastosowania bioinżynierskie modelu CHAID, jak np. ocena zanieczyszczenia fumonizyną w kukurydzy wykonana za pomocą spektroskopii w bliskiej podczerwieni [114].

Klasycznym modelem drzewa decyzyjnego jest model CART (Classification and Regression Trees) zaproponowany przez Briemana i in. w 1984 r. [115]. Przeszukuje on wszystkie możliwe podziały i wybiera optymalny, bazując na mierze niejednorodności węzła. Takie podejście wiąże się z długim czasem przeprowadzania obliczeń. Ponadto algorytm wybiera zmienne do podziału w sposób obciążony tj. wybiera te, które prowadzą do większej liczby podziałów. Przedstawione wady zostały wyeliminowane w algorytmie QUEST (Quick Unbiased Efficient Statistical Tree) zaproponowanym w 1997 przez Loha i Shiha [116]. Podział węzła podczas wykorzystania tej metody bazuje na kwadratowej analizie dyskryminacyjnej. QUEST wykorzystuje się np. podczas klasyfikacji odbiciowych widm średniej podczerwieni pochodzących od produktów spożywczych takich jak miód [117]. W artykule [118] autorzy przedstawili

różne modele drzew decyzyjnych pracujących na wynikach spektroskopii Ramanowskiej NIR, w zastosowaniu do diagnostyki raka żołądka.

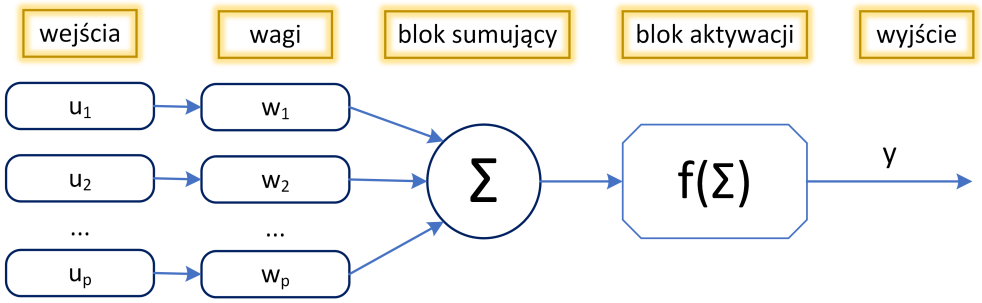
Drzewa klasyfikacyjne można wykorzystywać zarówno w celach przyporządkowania cech ilościowych jak i jakościowych. Ich klarowna konstrukcja pozwala na pełną analizę procesu decyzyjnego, a prosta forma końcowa na szybką i efektywną klasyfikację nowych obiektów [76]. Ponadto omawiane algorytmy są odporne na obserwacje odstające, jednak charakteryzują się wrażliwością na kolejność danych w zbiorze zarówno tych do nauki, jak i tych do weryfikacji wyniku [104]. Zastosowanie np. k-krotnej kroswalidacji, inaczej k-krotnego testu krzyżowego, umożliwia wielokrotne powtórzenie procedury uczenia na zmiennym zestawie danych uczących i weryfikacyjnych, co w konsekwencji niweluje problem wrażliwości klasyfikatora na kolejność danych w obu zbiorach (więcej informacji nt. tej metody znajduje się w podrozdziale 2.2.5). Wartość parametru k należy przyjąć w sposób eksperymentalny.

Sztuczne sieci neuronowe

Działanie mózgu, przepływające w nim sygnały czy umiejętność rozwiązywania problemów, były inspiracją do próby stworzenia połączonych ze sobą sztucznych komórek nerwowych wykorzystujących model matematyczny do przetworzenia informacji. Pierwszy formalny model neuronu powstał już 1943 r. [119]. Jego idea nadal stanowi podstawę działania większości wykorzystywanych modeli. Polega ona na wywołaniu funkcji aktywacji na sumie sygnałów wejściowych z odpowiednią wagą. Przełomowym odkryciem było stwierdzenie opisane regułą Hebba [120] mówiące, iż informacja może być przechowywana w strukturze połączeń między neuronami. Zaowocowało to stworzeniem metody uczenia sieci neuronowej polegającej na zmianach wag tychże połączeń.

Jedną z pierwszych działających sieci neuronowych był perceptron (zbudowany w 1958 r. przez Rosenblatta i Wightmana w Cornell Aeronautical Laboratory) – maszyna klasyfikująca obrazy i zdolna do uczenia się poprzez modyfikację połączeń prowadzących do układów progowych. W przypadku, gdy perceptron ma p wejść, wówczas dzieli p–wymiarową przestrzeń na dwie półprzestrzenie, które rozdzielone są $(p-1)$–wymiarową hiperpłaszyzną, zwaną granicą decyzyjną [76].

Sieci neuronowe wykorzystywane są do zadań predykcyjnych, klasyfikacyjnych i tych związanych ze sterowaniem. Stosuje się je niemalże we wszystkich dziedzinach, gdzie pojawiają się ww. zagadnienia. Ich głównymi zaletami są: mała moc obliczeniowa potrzebna do użycia sieci (ale nie do treningu) i prostota użycia. Sieci należą do bardzo zaawansowanych technik

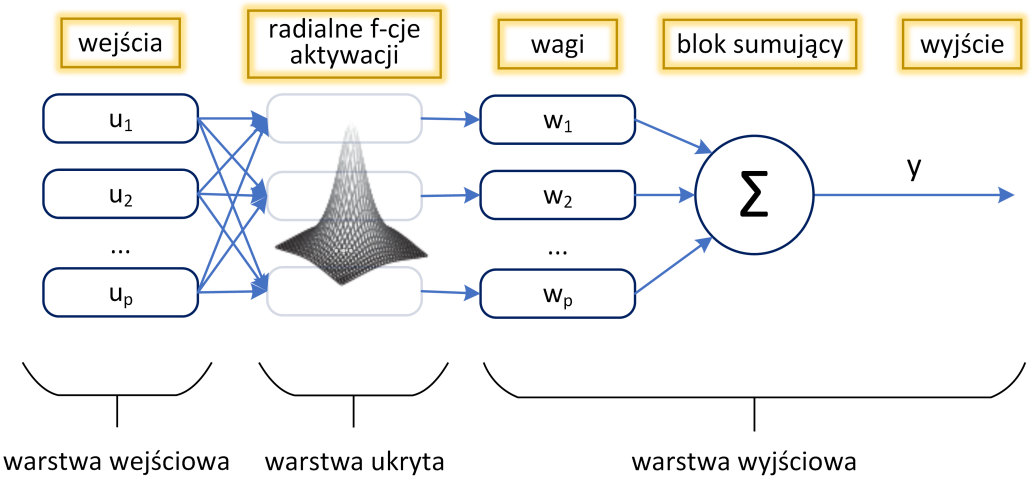


Rysunek 2.5: Graficzna wizualizacja modelu neuronu.

modelowania. Są w stanie odwzorować złożone funkcje, umożliwiają swobodne tworzenie modeli nieliniowych. Stosowanie sieci dobrze sprawdza się podczas modelowania funkcji nieliniowych w przypadku problemów wielowymiarowych, kiedy problemem staje się duża liczba zmiennych niezależnych. Prostota użytkowania sieci przejawia się w automatyzmie działania algorytmu, jednak nie oznacza wyeliminowanie roli operatora. W wariancie uczenia z nauczycielem powinien on przygotować dane stanowiące przykłady interesującej go zależności, wybrać właściwy rodzaj i konstrukcję sieci oraz zinterpretować wyniki. Sieć ma za zadanie nauczyć się reguł poprzez zmianę wag, bazując jedynie na przykładach podanych przez użytkownika.

Sieć neuronowa składa się z szeregu połączonych ze sobą neuronów. Pojedynczy neuron (schematycznie przedstawiony na rysunku 2.5) zwraca wartość funkcji aktywacji (y) wywołanej na sygnałach wejściowych $(u_1, u_2, ...u_p)$ przemnożonych przez swoje wagi $(w_1, w_2, ...w_p)$ i zazwyczaj zsumowanych do pojedynczego argumentu. Jeśli funkcją aktywacji jest funkcja liniowa, powstały neuron nazywany jest liniowym, analogicznie w przypadku neuronu nieliniowego, aktywowanego funkcją nieliniową. Nauka neuronu polega na dopasowywaniu (w trakcie procesu trenowania sieci) wag do konkretnych sygnałów wejściowych. To właśnie w macierzy wag zakodowana jest "cała wiedza" neuronu.

Podobnie, jak w szeroko pojętym uczeniu statystycznym, w dziedzinie sieci neuronowych wyróżnia się uczenie z nauczycielem i bez nauczyciela. Występują również znaczne analogie w rozumieniu ww. wariantów. Podczas nauki sieci z nauczycielem podawane są prawidłowe odpowiedzi na konkretne sygnały wejściowe i porównywane są z tymi na wyjściu sieci. Różnica jest traktowana jako błąd i generuje zjawisko uczenia się sieci. Najpopularniejsza strategia wykorzystująca uczenie z nauczycielem nazywana jest metodą wstecznej propagacji błędu. To ona była kolejnym przełomem podczas rozwoju dziedziny sztucznych sieci neuronowych.



Rysunek 2.6: Graficzna wizualizacja modelu sieci RBF.

W wariancie nauki sieci bez nauczyciela nie podaje się prawidłowej odpowiedzi, sieć sama musi wydobyć kategorie lub cechy charakterystyczne ze zbioru danych wejściowych.

Zasadę działania sieci neuronowych można podsumować stwierdzeniem, iż w procesie uczenia, sieć dąży do zminimalizowania różnicy między sygnałem na wyjściu neuronu, a odpowiedzią.

Sieci neuronowe wykorzystują wiele różnych algorytmów uczenia. W literaturze można znaleźć opisy takich typów sieci jak: sieci wielowarstwowe, rekurencyjne, samoorganizujące z konkurencją, rezonansowe, probabilistyczne sieci neuronowe czy sieci o radialnych funkcjach bazowych (RBF). Uczenie każdej sieci powinno być wykonywane na próbce reprezentatywnej. Oznacza to, że:

1. każda próbka powinna być losowo wybrana z poszczególnej klasy;
2. próbki powinny być dostatecznie liczne.

Spełnienie drugiego warunku nie jest jednoznaczne, ponieważ zależy od wielu czynników tj.: poziomu zaszumienia danych, stopnia nakładania się klas na siebie czy liczby cech opisujących obiekty. Można założyć słuszność stwierdzenia, że im próbka jest liczniejsza, tym lepiej odzwierciedla własności poszczególnych klas. W przypadku danych spektralnych może występować zarówno problem zaszumienia, liczności cech opisujących dane, jak również dostępności próbek (równoznaczne z ich małą liczbą), które to kwestie negatywnie oddziałują na końcowy wynik klasyfikacji uzyskany za pomocą sieci neuronowych.

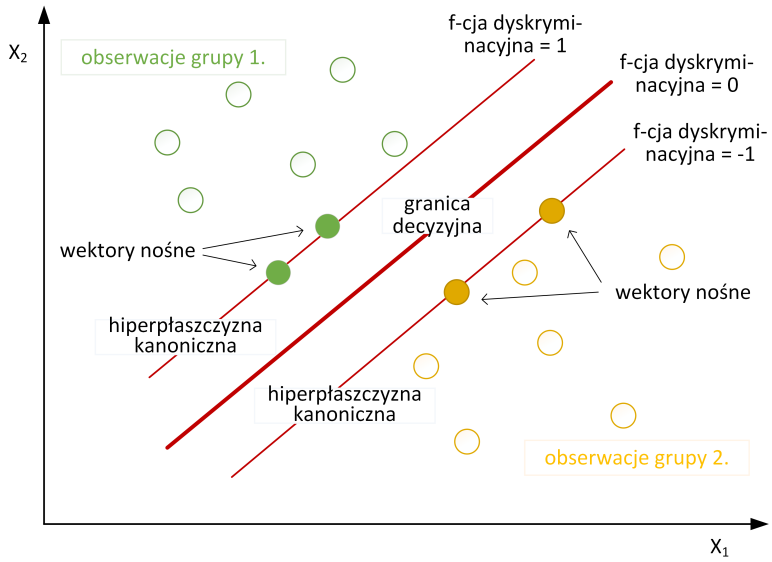
Modele RBF (rysunek 2.6) wykonują nieliniowe przekształcenia przestrzeni danych wejściowych za pomocą warstwy neuronów ukrytych reprezentujących zmienne funkcje radialne.

Neuron radialny, występujący w sieci RBF, reprezentuje hipersferę, która dokonuje podziału kołowego wokół punktu centralnego. RBF może służyć do nieliniowej aproksymacji funkcji, jak również z powodzeniem wykorzystuje się go w zagadnieniach klasyfikacyjnych. Ta sieć działa również, jako uniwersalny aproksymator i przybliża z dowolną dokładnością każdą funkcję ciągłą [76]. Twierdzenie Covera [121] o separowalności wzorców jest bazą działania algorytmu RBF. Omawiana sieć składa się z trzech warstw: warstwy wejściowej, warstwy ukrytej z neuronami o radialnych funkcjach aktywacji oraz warstwy wyjściowej zawierającej neurony liniowe. Taka struktura, zgodnie z twierdzeniem Covera, przy dostatecznie dużej liczbie neuronów w warstwie ukrytej, zapewnia rozwiązanie problemu klasyfikacji nieliniowej. Centra neuronów są inicjowane podczas tworzenia się struktury sieci. W czasie uczenia sieci, centra dostosowują się do danych uczących. Neurony liniowe z warstwy wyjściowej sumują sygnały pochodzące z warstwy ukrytej i przekazują jako wynik odpowiadający wektorowi wejściowemu, jednocześnie mają przypisany wektor wag i wartość progową.
Uczenie sieci RBF można przedstawić następująco:

1. dobór kształtu oraz położenia funkcji bazowych wykonywany na podstawie:
   - doboru losowego;
   - metody k-średnich;
   - metody wstecznej propagacji błędu;
2. dobór macierzy wag za pomocą metody pseudoinwersji macierzy Greena [76].

Sieci RBF przeważnie wymagają użycia większej liczby neuronów niż sieci jednokierunkowe, ale jednocześnie uczenie ich trwa krócej niż sieci perceptronowych. RBF wykorzystują aproksymację typu lokalnego, której zasięg działania jest bardzo ograniczony i skoncentrowany wokół centrów. Jest to powodem niższych możliwości uogólniania sieci RBF w stosunku do sigmoidalnych. Uważa się, że sieci radialne lepiej rozwiązują zadania klasyfikacyjne niż sieci sigmoidalne [76].

Szacuje się, że cały ludzki mózg może przetwarzać 10^{18} operacji logicznych na sekundę, podczas gdy 64–bitowy procesor PowerPC 970 wykonuje jedynie 10^{11} takich operacji [122]. Jednak należy zwrócić uwagę, że pojęcie operacji logicznej zachodzącej w mózgu jest sformowaniem mało precyzyjnym i trudnym do udowodnienia w sposób eksperymentalny.



Rysunek 2.7: Graficzna wizualizacja zasady działania SVM dla danych liniowo separowalnych.

Maszyna wektorów nośnych

Autorstwo idei maszyny wektorów nośnych (SVM) przypisuje się Vapnikowi i Chervonenkisowi (1971 r., 1974 r.) [123, 124]. SVM wykonuje ortogonalną transformację zamieniającą zbiór skorelowanych zmiennych na dane liniowo od siebie niezależne [125]. W przypadku rozważania zagadnienia dyskryminacji dwóch populacji liniowo separowalnych zadaniem maszyny wektorów nośnych jest wybranie takiej liniowej reguły klasyfikacyjnej, dla której odpowiednia hiperpłaszczyzna jest maksymalnie odległa od najbliższej jej obserwacji pochodzącej z próby uczącej (rysunek 2.7). Hiperpłaszczyzną kanoniczną nazywa się taką hiperpłaszczyznę, dla której moduł wartości funkcji dyskryminacyjnej dla próbki uczącej położonej najbliżej niej jest równy jedności. Wszystkie obserwacje ze zbioru danych uczących znajdujące się na hiperpłaszczyznach kanonicznych nazywa się wektorami nośnymi. Odpowiadają im niezerowe mnożniki Lagrange'a. O ostatecznej postaci funkcji dyskryminacyjnej decydują wyłącznie wektory nośne. Im większa wartość mnożnika Lagrange'a wektora nośnego, tym większy jest jego wpływ na kształt granic decyzyjnych. Usunięcie z próby dowolnej innej obserwacji nie wpłynie na postać hiperpłaszczyzny.

W praktyce, bardzo rzadko spotyka się dane liniowo separowalne. Oznacza to, że nie istnieje hiperpłaszczyzna rodzielająca klasy, która zapewnia poprawną klasyfikację wszystkich elementów zbioru uczącego. Zadaniem SVM, w takim przypadku, będzie wyznaczenie hiperpłaszczyzny minimalizującej prawdopodobieństwo błędnej klasyfikacji. Osiągane jest to poprzez transformację nieliniową elementów zbioru uczącego, z wyjściowej przestrzeni cech do

przestrzeni wyższego wymiaru (często jest to wymiar nieskończony), a następnie zastosowanie modelu liniowego w nowej przestrzeni. Wynikiem działania algorytmu SVM jest wektor wag oraz przesunięcie hiperpłaszczyzny kanonicznej opisywane funkcją dyskryminacyjną.

W przypadku zagadnienia klasyfikacji więcej niż dwóch klas, podstawową metodą przy użyciu SVM jest rozszerzenie opisanego powyżej modelu binarnego poprzez konstrukcję liczby funkcji dyskryminacyjnych odpowiadającej liczbie klas. Wszystkie funkcje dyskryminacyjne rozpatruje się równocześnie uogólniając funkcję straty (dobranie najlepszych hiperpłaszczyzn). Osłabiony zostaje warunek poboczny poprzez brak konieczności zerowania się funkcji dyskryminacyjnych na granicach decyzyjnych.

Pierwotnie miarę złożoności klasyfikatora utożsamiano z liczbą jego parametrów. Stwierdzenie to ewoluowało do określenia, że rodzina funkcji dyskryminacyjnych ma wymiar równy $p+1$, gdzie p jest maksymalną liczbą punktów liniowo niezależnych w przestrzeni zawierającej hiperpłaszczyzny. Im mniejszy wymiar klasyfikatora, tym mniejsza złożoność funkcji dyskryminacyjnej. Zatem dąży się do minimalizacji wymiaru np. poprzez maksymalizację odległości pomiędzy hiperpłaszczyznami kanonicznymi, co wiąże się również z minimalizacją rzeczywistego błędu przyporządkowania. Równowagę pomiędzy dopuszczeniem błędnej klasyfikacji pewnych elementów próby uczącej, a złożonością klasyfikatora uzyskuje się przy zastosowaniu np. kroswalidacji, która jest również wykorzystywana podczas testowania takich modeli jak DT. Ostateczne rozwiązanie zadania optymalizacyjnego nie zależy od obserwacji nie będących wektorami nośnymi. Jeśli są one poprawnie klasyfikowane przez klasyfikator zbudowany na pełnej próbie, będą również poprawnie przyporządkowane podczas kroswalidacji. Oznacza to, że SVM cechuje się dobrymi własnościami reguł klasyfikacyjnych o stosunkowo małej liczbie wektorów nośnych. Własność ta jest tym bardziej znacząca w zastosowaniu przestrzeni wielowymiarowych. Potwierdza się to w licznych publikacjach z tematyki pomiarów spektralnych wykorzystujących SVM jak np. klasyfikacja obrazów spektralnych na podstawie nielicznej próby odpowiedzi widmowych pochodzących ze wskazanych pikseli zarejestrowanych obrazów [48]. SVM jest często wykorzystywaną techniką w szeroko pojętym zastosowaniu spektroskopii IR [126].

Czas wykonania algorytmu SVM zależy przede wszystkim od liczby obserwacji ciągu uczącego. Użytkownik może regulować parametr mający wpływ na osiągnięcie optymalnego stosunku pomiędzy złożonością klasyfikatora, a oceną aktualnego poziomu błędu uzyskaną metodą resubstytucji. Bardziej skomplikowanie przedstawia się sytuacja wyboru kryterium

funkcji jądra, o którym również decyduje operator. Nawet w przypadku ustalenia typu jądra zachodzi potrzeba estymacji jego parametrów, która wykonywana jest najczęściej empirycznie poprzez minimalizację poziom błędu na próbie testowej lub ewentualnie na danych uczących np. za pomocą kroswalidacji. Nieumiejętne wykonanie tych operacji grozi nadmiernym dopasowaniem klasyfikatora do danych uczących. Interpretacja wyników SVM nie jest intuicyjna, co powoduje, że metoda ta często wykorzystywana jest na zasadzie "czarnej skrzynki".

2.2.5. Miary oceny jakości klasyfikacji

Prawidłowa klasyfikacja polega na maksymalizacji wartości stopnia przyporządkowania obiektów badanych do specyficznych klas. Ma to nastąpić przy jednoczesnej generalizacji klasyfikatora, czyli nie dopuszczeniu do nadmiernego dopasowania się modelu do próbek zbioru uczącego. W literaturze występują różne metody oceny generalizacji algorytmów. Dobierane są one pod względem czasu uczenia, rozmiaru pliku uczącego oraz od tego, czy występuje plik testowy.

Wskazane jest wykorzystanie zbioru danych testowych, ponieważ umożliwia on wiarygodne sprawdzenie badanego klasyfikatora pod względem możliwości uogólniających. Danymi testowymi nazywa się obiekty nie wchodzące w skład zbioru uczącego, czyli próbki nowe dla klasyfikatora. W przypadku braku wyodrębnienia z danych wejściowych zbioru testowego przeprowadza się ocenę klasyfikatora na podstawie losowego podziału danych na treningowe i testowe. Podstawowymi metodami tego typu są: kroswalidacja (inaczej test krzyżowy) [127], kroswalidacja z wykorzystaniem wyboru cech za pomocą metody Monte Carlo [128], czy i bootstrapping [129].

Bazową techniką przeprowadzenia kroswalidacji jest walidacja prosta. Początkowy zbiór danych jest dzielony w sposób losowy na dwa rozdzielne zbiory: uczący (zawierający ok 70% wszystkich próbek) i testowy (zawierający pozostałe próbki). Do stworzenia modelu klasyfikator wykorzystuje jedynie dane będące w zbiorze uczącym. Następnie do tak wygenerowanego modelu wprowadzane są dane testowe i dokonywana jest predykcja klas zakończona obliczeniem współczynnika dokładności, który w rozprawie oznaczony jako $acc\ r$.

Najbardziej rozpowszechnionym sposobem sprawdzianu krzyżowego jest k-krotna kroswalidacja. Zasadę jej działania można przedstawić następująco:

1. losowy podział zbioru danych na k części, o numerach od 1 do k;

2. nauka klasyfikatora na danych składających się ze zbiorów od 1 do $k-1$;
3. testowanie tak nauczonego klasyfikatora na zbiorze o numerze k (zbiór testowy);
4. obliczenie błędu tej klasyfikacji (stosunek błędnie sklasyfikowanych próbek zbioru testowego do liczebności tego zbioru);
5. wykonanie pkt 2. – 4. k razy, każdorazowo wybierając kolejny zbiór, który najpierw jest usuwany z danych uczących, następnie staje się danymi testowymi;
6. obliczenie średniej arytmetycznej poszczególnych błędów klasyfikacji.

Wynik uzyskany z pkt 6. nazywany jest błędem kroswalidacji. Dzięki takiemu podejściu, każda próbka występuje tylko raz w zbiorze testowym i $k-1$ razy w zbiorze uczącym.

Kroswalidacja Monte Carlo, działa na zbliżonej zasadzie, z tą różnicą, że podział na dane uczące i testowe zazwyczaj występuje w stosunku 2 do 3 i wykonywany jest w sposób losowy. Proces generowania zbiorów jest każdorazowo powtarzany na nowo (najczęściej 30-50 razy) [130]. Podobnie jak podczas walidacji prostej wynikiem końcowym jest średnia wartości cząstkowych. Czynnikiem odróżniającym te dwa podejścia jest fakt, że podczas kroswalidacji Monte Carlo kolejne podziały nie opierają się na rozłącznych podzbiorach. Oznacza to, że konkretne obserwacje mogą znaleźć się wielokrotnie w zbiorze uczącym, jak również w zbiorze testowym.

Metoda bootstrap wykorzystuje cechy k-krotnej kroswalidacji i Monte Carlo. Wykonywane jest k-krotne losowanie zbioru n-elementowego ze zwracaniem, w ten sposób generowane jest k zbiorów uczących. Na pozostałych danych wykonywane jest testowanie klasyfikatora. Wynikiem metody jest średnia arytmetyczna błędów obliczonych przy każdorazowym losowaniu.

Na małych zbiorach danych można stosować kroswalidację typu *leave-one-out*, gdzie w zbiorze testowym jest tylko jedna próbka, jednak jest to bardzo obciążający obliczeniowo proces i znacznie wydłuża procedurę testowania.

Ze względu na losowanie zbiorów uczących i testowych, podczas badań należy wielokrotnie wykonać obliczenia oceniające zadany klasyfikator.

Maksymalizując wartości dokładności przyporządkowania błąd klasyfikatora jest minimalizowany. Łączny błąd klasyfikacji określa się jako stosunek błędnie sklasyfikowanych próbek zbioru testowego do liczebności tego zbioru. Jego dopełnieniem jest współczynnik dokładności – acc *r.*, definiowany jako stosunek poprawnie sklasyfikowanych próbek zbioru testowego do liczebności tego zbioru. W przypadku korzystania z DT w połączeniu z użyciem testu krzyżowego, dokładność (w rozprawie oznaczona – acc) można przedstawić jako dopełnienie

błędu klasyfikacji DT do jedności. Błędem klasyfikacji DT nazywa się iloczyn błędu w węźle głównym i błędu kroswalidacji. Błąd w węźle głównym definiuje się jako procent błędnie sklasyfikowanych próbek w węźle głównym [131].

W części eksperymentalnej rozprawy podstawowym wskaźnikiem jakości klasyfikatora jest dokładność DT – acc – w przypadku wykorzystania drzewa decyzyjnego z kroswalidacją oraz współczynnik dokładności modelu – $acc\ r.$ – dla pozostałych klasyfikatorów (np. bazujących na ANN i SVM) i przy ponownym sprawdzeniu klasyfikatora DT.

2.3. Podsumowanie rozdziału

W niniejszym rozdziale przedstawiono ogólne podejście do procesu klasyfikacji wykorzystującego pomiary spektralne, wyszczególniając zagadnienie wstępnego przetwarzania danych, selekcję cech sygnałów widmowych, wykorzystanie klasyfikatora oraz ocenę końcowego przyporządkowania. Przedstawiono sposoby eliminacji wpływu na wynik końcowy charakterystyki podłoża, na którym znajduje się badany materiał oraz charakterystyki wykorzystywanego źródła światła. Dzięki tym operacjom matematycznym uzyskuje się transmitancję próbki, którą można poddać procesowi filtracji. Następnie wskazane jest, aby zredukować wymiarowość tak przygotowanego sygnału. Po selekcji cech następuje wykorzystanie klasyfikatora i końcowe przyporządkowanie do odpowiedniej klasy. Ważnym etapem procesu klasyfikacji jest jej ocena. Dzięki niej możliwe jest określenie, czy klasyfikator nie jest nazbyt dopasowany do danych. Najczęściej jest to wykonywane za pomocą odpowiednich procedur operujących na podzielonym zbiorze danych na zbiór uczący i testowy.

Każdy z wymienionych etapów można wykonać na różne sposoby, wykorzystując w tym celu mniej bądź bardziej popularne sposoby. W niniejszym rozdziale zostały opisane przykładowe metody będące składowymi poszczególnych etapów.

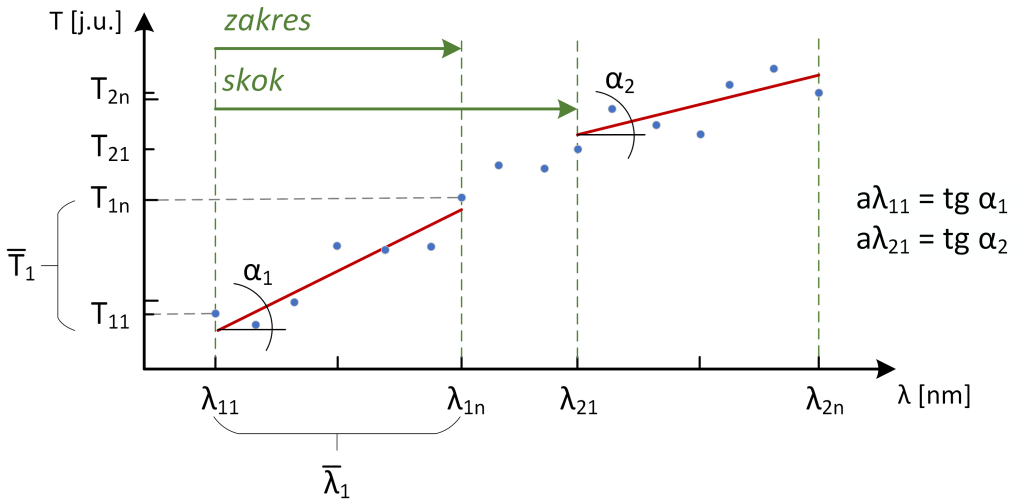
3. Metoda parametryzacji i metoda redukcji sygnału

3.1. Idea działania proponowanych metod

Proponowane w rozprawie metody łączą w sobie nowatorskie podejście do procesu redukcji wymiarowości oraz wykorzystanie znanego klasyfikatora w postaci drzewa decyzyjnego. Metoda parametryzacji z użyciem aproksymacji wielomianowej (metoda PAW) oraz metoda doboru i redukcji widma (metoda DRW) zmniejszają wymiarowość zakwizycjonowanych danych. Efektem pracy pierwszej z nich jest wygenerowanie dwukolumnowych macierzy zawierających nowo obliczone atrybuty sygnału badanego materiału wraz z odpowiadającymi im zakresami spektralnymi. Wykorzystanie drugiej metody pozwala na usunięcie wyselekcjonowanych fragmentów, które pogarszają lub nie zmieniają wyników ostatecznej klasyfikacji. Wycięte dane odpowiadają konkretnym, ustalonym przez użytkownika zakresom. Nie są one wąskimi, rozrzuconymi po pełnym widmie, obszarami spektralnymi trudnymi do interpretacji fizycznej.

Rozpoczęcie analizy obiektu będącej pierwszym etapem prezentowanego procesu klasyfikacji jest akwizycja sygnału spektralnego. Następnie, w zależności od wykorzystanego układu optycznego, obliczana jest transmitancja lub reflektancja badanego materiału. Uzyskany w ten sposób, uniezależniony od bazy i źródła światła sygnał, poddawany jest odszumieniu za pomocą filtra S-G, a następnie mogą zostać przeprowadzone na nim dalsze procedury metod PAW oraz DRW. W dalszej części opisu działania obu metod będzie wykorzystywane sformułowanie transmitancja, lecz analogiczne rozważania można przeprowadzić dla reflektancji.

Kluczowym etapem działania metody PAW jest wykonanie aproksymacji wielomianowej metodą najmniejszych kwadratów dla wielomianów stopnia co najwyżej pierwszego w konkretnych zakresach spektralnych transmitancji określonych przez parametry wskazane przez użytkownika. Operator definiuje zakresy, bądź wartości dwóch parametrów: $zakres$ i $skok$.



Rysunek 3.1: Graficzna wizualizacja parametrów: $zakres$ i $skok$ wraz z aproksymacją wielomianów stopnia 1. metodą najmniejszych kwadratów. Tangensy kątów nachylenia prostych zaznaczonych na czerwono są ich współczynnikami kierunkowymi oznaczonymi $a\lambda_{11}$ i $a\lambda_{21}$.

$Zakres$ określa przedział widma, z którego dane są podstawą do wyliczenia wielomianu tego fragmentu sygnału. $Skok$ jest wartością oznaczającą odległość pomiędzy początkiem jednego i początkiem kolejnego dopasowania. Oba parametry podawane są w nm. Rysunek 3.1 przedstawia graficzną interpretację parametrów $zakres$ i $skok$, dopasowane proste metodą najmniejszych kwadratów oraz oznaczenia, które posłużą do matematycznego opisu wyznaczania parametrów a – kierunkowych prostych – będących częściowym efektem działania metody PAW.

Aproksymacja wielomianowa metodą najmniejszych kwadratów dla wielomianów stopnia co najwyżej pierwszego ma za zadanie wyznaczyć takie proste umiejscowione we wskazanych $zakresach$, których sumy (S) odległości poszczególnych punktów pomiarowych od prostych będą minimalizowane (wzór 3.1). Opis matematyczny za pomocą wzorów 3.1, 3.2, 3.3 ogranicza się do analizy przypadku dopasowania prostej z indeksem 1 przedstawionej na rysunku 3.1.

$$S_1 = \sum_{i=1}^{n} [T_{1i} - T(\lambda_{1i})]^2 = \sum_{i=1}^{n} (T_{1i} - a\lambda_{1i} - b)^2 \tag{3.1}$$

Gdzie:

n – liczba punktów pomiarowych zawierających się w $zakresie$;

i – indeksy punktów pomiarowych, wartości od 1 do n;

b – współczynnik przesunięcia prostej po osi y, nie brany pod uwagę w dalszych rozważaniach.

Współczynniki kierunkowe prostych a, których wartości są podstawą do dalszej pracy algorytmu PAW, oblicza się wykonując x razy obliczenia opisane wzorem 3.2 lub 3.3, gdzie x oznacza liczbę *skoków* zawartych w badanym przedziale widmowym transmitancji.

$$a\lambda_{11} = \frac{n\sum_{i=1}^{n}\lambda_{1i}T_{1i} - \sum_{i=1}^{n}\lambda_{1i}\sum_{i=1}^{n}T_{1i}}{n\sum_{i=1}^{n}\lambda_{1i}^2 - (\sum_{i=1}^{n}\lambda_{1i})^2} \tag{3.2}$$

Co jest równoznaczne:

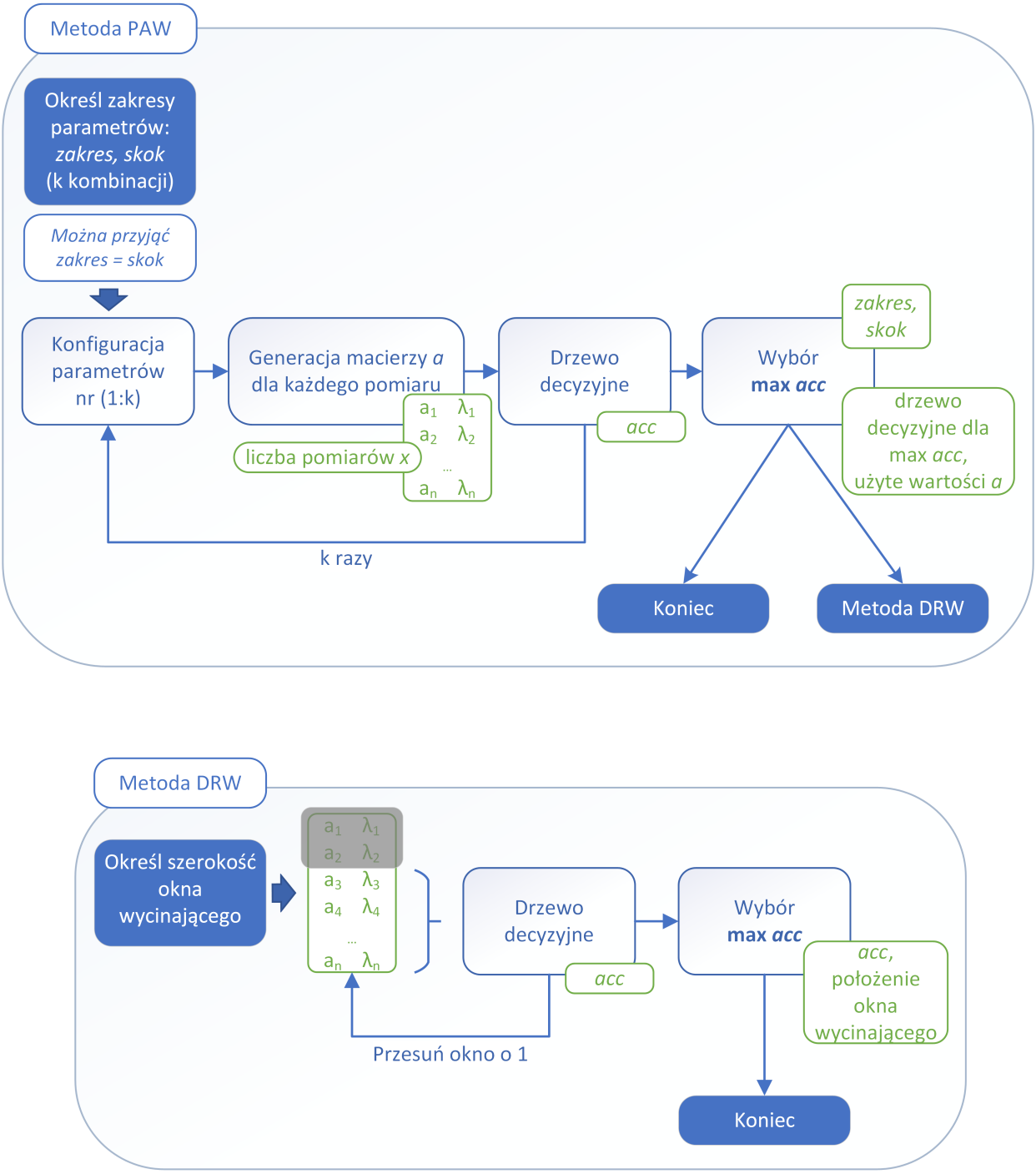
$$a\lambda_{11} = \frac{\sum_{i=1}^{n}\lambda_{1i}T_{1i} - n\overline{\lambda}_1\overline{T}_1}{\sum_{i=1}^{n}\lambda_{1i}^2 - n(\overline{\lambda}_1)^2} \tag{3.3}$$

gdzie:

$\overline{\lambda}_1$ – wartość średnia długości fal z przedziału pierwszego;
$\overline{T}_1$ – wartość średnia transmitancji z przedziału pierwszego.

Dla każdego pomiaru zostaje wygenerowana dwuwymiarowa macierz cech, ozn. macierz a, której jednym wymiarem są wartości współczynników kierunkowych, ozn. a, dopasowanych prostych metodą najmniejszych kwadratów do fragmentów danych, a drugim – odpowiadające współczynnikom a wartości długości fal. Proste są wielomianami stopnia co najwyżej pierwszego dopasowanymi do fragmentów widm transmitancji. Wartość a liczbowo określa nachylenie prostej.

Metoda PAW, przedstawiona w sposób schematyczny na rysunku 3.2, wykorzystuje klasyfikator DT. Dla konkretnej kombinacji parametrów, czyli dla zbioru dwukolumnowych macierzy a o liczebności równej liczbie pomiarów, wykonywany jest algorytm DT wraz z 7–krotną kroswalidacją. Wynikiem jest wartość acc i model drzewa decyzyjnego. Po wykonaniu k–krotnie (gdzie k oznacza numer kombinacji parametrów) powyższej procedury następuje wybór maksymalnej wartości acc w celu znalezienia najbardziej efektywnego zestawu wartości: $zakres$ i $skok$. Jest to jednoznaczne ze wskazaniem modelu drzewa, którego efektem jest najwyższe acc, co za tym idzie, wybrany zbiór parametrów a. Wskazanie parametrów a, będących jednoznacznie przypisanych do odpowiadającym im zakresom długości fal, jest ważne ze względu na przeprowadzenie ewentualnych późniejszych procedur. Na tym etapie proces może zostać zakończony lub użytkownik może zdecydować o rozpoczęciu wykonania metody DRW.



Rysunek 3.2: Schemat metody PAW – parametryzacji z użyciem aproksymacji wielomianowej oraz metody DRW – doboru i redukcji widma. Zielona ramka – wynik konkretnego etapu, szary prostokąt – okno wycinające o zadanej szerokości.

Algorytm DRW tworzy okno wycinające fragment danych (na rysunku 3.2 przedstawiony jako szary prostokąt) o szerokości zadanej przez użytkownika i ustawia je z brzegu zakresu spektralnego w ten sam sposób na wszystkich macierzach a odpowiadającym przeprowadzonym pomiarom. Szerokość okna jest na tyle duża, aby po usunięciu danych „zakrywanych" przez konkretne okno, można było zinterpretować je fizycznie. W tym przypadku oznacza to np. możliwość dopasowania elementów optycznych poprzez dodanie konkretnego filtra spektralnego ograniczającego wiązkę, lub wykorzystanie diod elektroluminescencyjnych o określonych barwach, jako źródła światła. W pierwszym położeniu okna następuje wykonanie obliczeń klasyfikacyjnych przy użyciu DT wraz z wykonaniem kroswalidacji i zapisanie wartości acc. Następnie okno zostaje przesunięte o 1 pozycję w stronę przeciwnego (w stosunku do miejsca początkowego położenia okna) krańca spektrum i procedura się powtarza do czasu, gdy okno osiągnie pozycję, w której usuwa ostatni wiersz macierzy a. Następuje wskazanie długości fali, od której usuwając dane spektralne o zakresie równym szerokości okna, osiągnięto maksymalne acc. Opisany proces jest powtarzany tyle razy, ile szerokości okien chce zweryfikować użytkownik.

Przykłady wykorzystania systemu z zaimplementowaną metodą PAW lub połączeniem metod PAW i DRW w rzeczywistych zastosowaniach przedstawione są w rozdziale 4.

3.2. Analiza

W celu zapoznania czytelnika bliżej z przedstawionymi w rozprawie metodami PAW i DRW, kolejne dwa podrozdziały poświęcone są opisowi zależności jakie wiążą poszczególne parametry między sobą i ich wpływem na wynik końcowy w postaci acc, czyli dokładnością klasyfikacji (szczegółowy opis wskaźników jakości klasyfikatorów, z uwzględnieniem acc znajduje się w rozdziale 2.2.5). Analiza algorytmów pozwalająca na zapoznanie się z relacjami wiążącymi poszczególne parametry została przeprowadzona na danych dotyczących klasyfikacji pochodzenia botanicznego próbek czterech różnych miodów. Dokładny opis badania z uwzględnieniem charakterystyki mierzonych materiałów oraz wykorzystanego układu optycznego, przedstawiony jest w rozdziale 4.2.2.

Podrozdział 3.2.1 zawiera informacje na temat zależności pomiędzy parametrami $zakres$ i $skok$ w kontekście tworzenia macierzy a.

Macierze a, zawierające informacje o parametrach a wraz z przypisanymi im odpowiednimi długościami fal są podstawą rozważań zaprezentowanych w podrozdziale 3.2.2. Opisana jest

w nim również procedura doboru takiego obszaru macierzy a, po usunięciu którego, uzyskuje się wzrost końcowego acc.

Podczas wykonywania wszystkich analiz wykorzystano drzewo decyzyjne o 7-krotnej kroswalidacji jako klasyfikator. Inni autorzy często przyjmują 10-krotny sprawdzian krzyżowy podczas testowania swoich modeli [132], choć również zalecane jest przyjęcie $k = 5$ [37]. Chcąc zoptymalizować poziom wariancji wyników oraz czas obliczeń, autorka zdecydowała się na wykorzystanie $k = 7$ podczas kroswalidacji.

Dane zostały odszumione za pomocą filtra S-G o szerokości okna dopasowania wielomianu równej 401 punktów pomiarowych. Wartość dobrano eksperymentalnie.

3.2.1. Metoda parametryzacji z użyciem aproksymacji wielomianowej – PAW

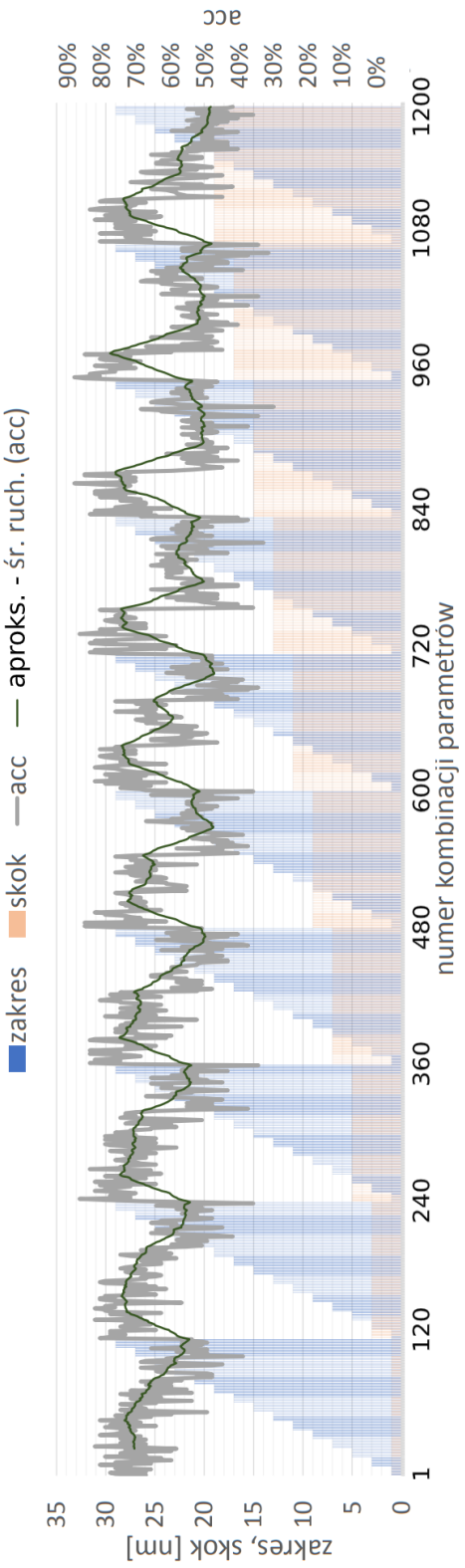
Podczas parametryzacji sygnału transmitancji lub reflektancji próbki, głównymi czynnikami oddziałującymi na efekt końcowy (dokładność klasyfikacji DT acc) są: $zakres$ i $skok$. Przeprowadzono analizę parametrów zawierających się w następujących przedziałach: $zakres$ – od 1 nm do 29 nm, co 2 nm oraz $skok$ – od 1 nm do 19 nm, co 2 nm. Przedziały wartości $zakres$ oraz $skok$ zostały dobrane ze względu na możliwość analizy trzech przypadków położenia wygenerowanych wielomianów co najwyżej stopnia pierwszego (prostych) względem siebie:

- proste częściowo na siebie nachodzą: $zakres > skok$;
- proste są od siebie oddalone: $zakres < skok$;
- w końcu jednej prostej znajduje się początek drugiej: $zakres$ = $skok$.

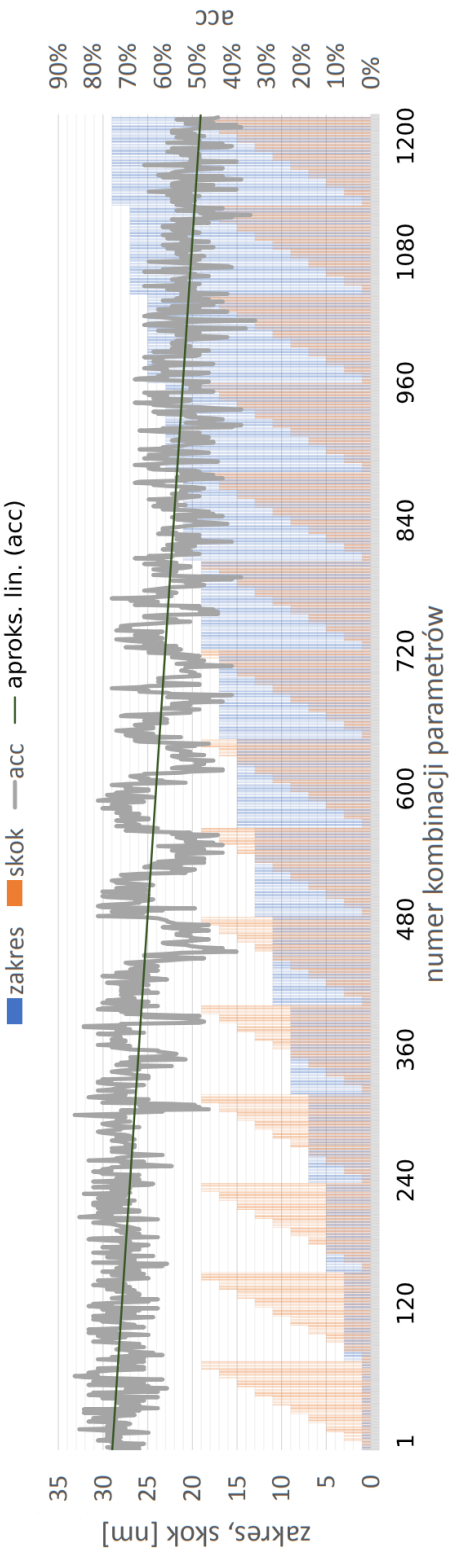
Obliczono wynikowe acc każdej kombinacji parametrów.

Wykres 3.3 przedstawia zależności pomiędzy wynikowym acc (bezpośrednie dane – szara linia, wygładzone za pomocą średniej ruchomej – czarna linia), a $zakresem$ (niebieskie słupki), w przypadku występowania stałego, zdyskretyzowanego przyrostu $skoku$ (czerwone słupki). Można zauważyć, że wzrost $skoku$ w minimalnym stopniu wpływa na tendencje zmian obserwowane na wykresie. Widoczne oscylacje acc wyraźnie korespondują z wartościami $zakresu$.

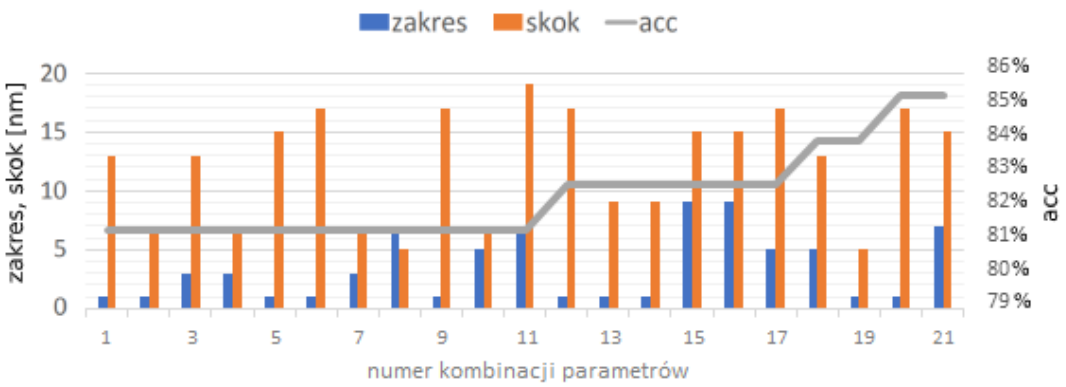
Relację acc ze stałym, zdyskretyzowanym wzrostem $zakresu$, przy jednoczesnym uwzględnieniu zmian $skoku$ przedstawia wykres 3.4. Zauważalna jest tendencja spadkowa wartości acc podczas wzrostu $zakresu$. Oznacza to, że średnio najwyższe wartości acc uzyskuje się dla wartości a będących efektem dopasowania prostych do wąskich zakresów spektralnych. Oscylacje szarej linii osiągają największe amplitudy w przedziale 11 – 19 nm parametru



Rysunek 3.3: Wartości dokładności acc w zależności od kombinacji parametrów $zakres$ i $skok$, z uwzględnieniem uszeregowania $skoku$ w sposób rosnący. Szara linia – acc [%]; zielona linia – aproksymacja za pomocą średniej ruchomej acc [%]; niebieskie słupki – $zakres$ [nm]; pomarańczowe słupki – $skok$ [nm]. Wykres należy czytać w następujący sposób: dla $zakresu$ x nm i $skoku$ y nm wartość acc wynosi z%. Oś x przedstawia kolejne kombinacje parametrów $zakres$ i $skok$.



Rysunek 3.4: Wartości dokładności acc w zależności od kombinacji parametrów $zakres$ i $skok$, z uwzględnieniem uszeregowania $zakresu$ w sposób rosnący. Szara linia – acc [%]; zielona linia – aproksymacja liniowa acc [%]; niebieskie słupki – $zakres$ [nm]; pomarańczowe słupki – $skok$ [nm]. Wykres należy czytać w następujący sposób: dla $zakresu$ x nm i $skoku$ y nm wartość acc wynosi z%. Oś x przedstawia kolejne kombinacje parametrów $zakres$ i $skok$.

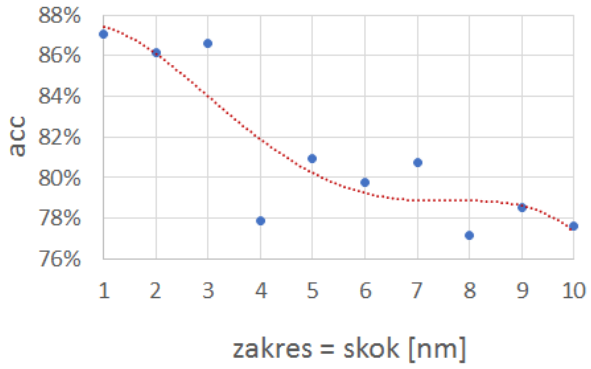


Rysunek 3.5: Kombinacje parametrów $zakres$ i $skok$ w wyniku zastosowania których uzyskuje się dokładności acc powyżej 80%. Szara linia – acc [%]; niebieskie słupki – $zakres$ [nm]; pomarańczowe słupki – $skok$ [nm]. Wykres należy czytać w następujący sposób: dla $zakresu$ x nm i $skoku$ y nm wartość acc wynosi z%. Oś x przedstawia kolejne kombinacje parametrów $zakres$ i $skok$.

$zakres$, wskazując mocniejsze oddziaływanie $skoku$ na acc w tym zakresie. Na pozostałych obszarach wpływ $skoku$ na acc jest pomijalny.

Analizując kombinację parametrów $zakres$ i $skok$, która doprowadziła do osiągnięcia acc większego od 80% (rysunek 3.5) można stwierdzić, że w przeważającej ilości przypadków, wartość $zakresu$ jest mniejsza niż wartość $skoku$. Nie powoduje to nakładania się na siebie prostych, czyli nie występuje overlapping danych (pojedyncza wartość z wykresu tylko raz jest uwzględniana w obliczeniach wielomianowych). Przy znacznej różnicy wartości $zakresu$ i $skoku$ (gdy $zakres < skok$) powstają obszary nieuwzględniane podczas parametryzacji. Wysokie wartości acc przy takich kombinacjach parametrów mogą oznaczać, że część danych jest zbędna w konkretnym procesie kwalifikacji lub nawet zaburzają końcowy wynik. Jest to wskazówka do wykonania dalszej redukcji macierzy a, poprzez wycięcie jej fragmentu, np. za pomocą metody DRW (rozdział 3.2.2).

Wykres 3.6 przedstawia rozkład uśrednionego z 50-ciu powtórzeń acc wyliczonego na podstawie sparametryzowanych wszystkich danych z całego zakresu widmowego. Przypadek zrównania się parametrów $zakres$ i $skok$ pokazany na wykresie oznacza wzięcie pod uwagę każdej danej wartości transmitancji tylko raz podczas procesu parametryzacji. Dla każdej pary $zakres$ = $skok$ obliczono macierz a, a następnie acc. Z wykresu, można w sposób jednoznaczny odczytać, że najwyższy poziom prawidłowej klasyfikacji przypada najmniejszej wartości $zakresu$, czyli 1 nm (rozdzielczość widmowa pomiaru wynosiła 0.1 nm). Macierz



Rysunek 3.6: Zależność uśrednionego z 50-ciu powtórzeń acc obliczonego na podstawie sparametryzowanych danych z pełnego zakresu spektralnego, w stosunku do kombinacji parametrów o zrównanych wartościach względem siebie. Czerwona kropkowana linia – aproksymacja wielomianowa wartości acc.

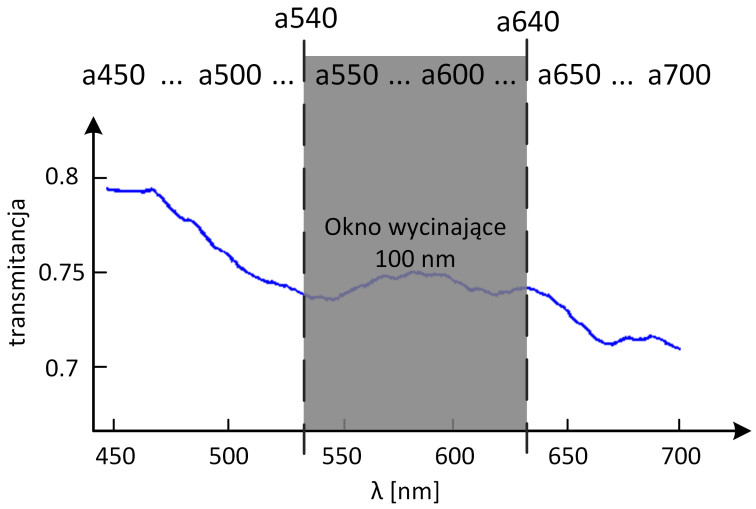
a obliczona na podstawie parametrów $zakres$ = $skok$ = 1 nm stała się podstawą do dalszego zmniejszania wymiarowości danych za pomocą metody DRW.

3.2.2. Metoda metoda doboru i redukcji widma – DRW

Testy metody redukcji, zostały wykonane na dwuwymiarowych macierzach a zawierających długości fal i odpowiadające im wartości a, będące efektem dopasowania prostych powstałych po parametryzacji wykresu transmitancji: $zakres$ = 1nm i $skok$ = 1nm. Każdemu pomiarowi odpowiadała jedna macierz a. Przeanalizowano okna wycinające o szerokościach: 10, 20, 40, 60, 100, 150, 200, 250 nm.

Poszczególne macierze poddano następującej procedurze:

1. umieszczenie pierwszego okna wycinającego na skraju zakresu spektralnego wykresu trasmitancji lub reflektancji;
2. obliczenie acc za pomocą drzewa decyzyjnego z 7-krotną kroswalidacją na podstawie danych pozostałych po wycięciu;
3. pięciokrotne powtórzenie działań z pkt 1 i 2 (w celu zmniejszenia rozrzutu wartości acc spowodowanego losowym doborem próbek podczas kroswalidacji), po czym zapisanie wyniku uśrednionego acc z pięciu powtórzeń;
4. przesunięcie okna wycinającego o 1 nm w stronę długich fal;
5. powtórzenie zadań z pkt 2, 3 (uwzględniając nowe położenie okna wycinającego) i pkt 4 do momentu takiego usytuowania okna, w którym zostanie wycięta największa wartość



Rysunek 3.7: Schematyczne przedstawienie idei działania okna wycinającego na danych będących transmitancją sygnału. Tu: okno o szerokości 100 nm, zakres 540–640 nm jest usuwany.

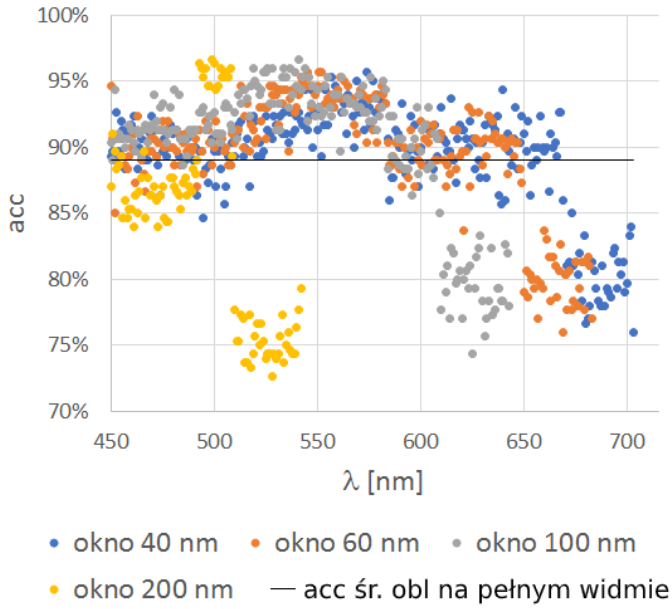
długości fali – w ten sposób zostanie przetestowany cały zakres danych pod względem częściowego ich usunięcia;

6. określenie szerokości okna oraz długości fali, której odpowiadające acc przyjmuje najwyższą wartość (wskazany punkt wyznacza długość fali, od której należy rozpocząć redukcję danych w zakresie wskazanym przez dobraną szerokość okna wycinającego);
7. stukrotne powtórzenie obliczeń acc dla wybranego położenia okna wycinającego i obliczenie średniego acc zredukowanych danych (zminimalizowanie wpływu wartości odstających z rozkładu wyników acc).

Ideę działania okna wycinającego pokazuje rysunek 3.7. W przedstawionym przypadku wynikowe acc obliczone jest na podstawie parametrów a, zawierających się w zakresach transmitancji pozostałych po usunięciu danych za pomocą okna wycinającego o szerokości 100 nm.

Wykres 3.8 przedstawia wartości acc obliczone po usunięciu odpowiednio umiejscowionych okien o wybranych szerokościach: 40, 60, 100 i 200 nm. Przedział wartości szerokości okien dobrano pod względem możliwości implementacji eksperymentalnej. Linia na poziomie 89% przedstawia średnie acc ze stu powtórzeń obliczone na danych pochodzących z pełnego zakresu spektralnego – bez wykorzystania okna wycinającego. Implementacja poszczególnych szerokości okien na długościach fal o największych wartościach z badanego przedziału cechuje się charakterystycznych spadkiem acc. Świadczy to o znaczącym wpływie na jakość klasyfikacji

danych będących z prawego krańca wykresu transmitancji. Innymi słowy, wycięcie wartości a opisujących prawą część wykresu transmitancji powoduje spadek skuteczności klasyfikacji. Największa wartość acc osiągana jest dla okna wycinającego o szerokości 100 nm począwszy od wartości 541 nm. Dla niniejszych wartości wykonano stukrotne powtórzenie obliczeń, których efektem było uzyskanie średniego acc na poziomie 95%. Wynik uległ poprawie o 6 punktów procentowych, w stosunku do analogicznych obliczeń na pełnym widmie.



Rysunek 3.8: Wartości dokładności acc obliczone po usunięciu odpowiednio umiejscowionych okien o szerokościach: 40 nm – niebieskie kropki, 60 nm – pomarańczowe kropki, 100 nm – szare kropki i 200 nm – żółte kropki, w zależności od położenia okna wycinającego na wykresie transmitancji (por. 3.7) (wartość długości fali odpowiada początkowej pozycji okna). Czarna linia – uśrednione ze stu powtórzeń acc obliczone na pełnym zakresie widma.

3.3. Wpływ niedokładności i rozdzielczości pomiaru urządzenia na działanie metod

Każdorazowy pomiar nie jest doskonały, zjawisko to opisywane jest za pomocą tzw. błędów pomiarowych. Są one sumą składowych przypadkowych i składowych systematycznych.

Błąd przypadkowy wynika ze stochastycznych czasowych i przestrzennych zmian wpływających na pomiar. Nie może on zostać skompensowany, jednak zwiększenie liczby pomiarów wykonanych w tym samym punkcie obserwacji i w ten sam sposób, a następnie wyciągnięcie z nich średniej kwadratowej spowoduje, że wartość oczekiwana błędu przypadkowego będzie dążyć do zera.

Błąd systematyczny jest efektem rozpoznanego działania wielkości wpływającej na wynik pomiaru, może być określony ilościowo. W przypadku występowania znacznego błędu systematycznego w porównaniu z wymaganą dokładnością pomiaru, należy go skompensować poprzez wprowadzenie addytywnie poprawki lub multyplikatywnie współczynnika poprawkowego. Oczekuje się, że po kompensacji wartość oczekiwana błędu będzie wynosić zero [133].

Po wykonaniu pomiaru i uwzględnieniu błędu pomiarowego, nie należy jednak wyniku uznawać za wartość prawdziwą, stanowiącą przedmiot badań - uznaje się, że takowa z zasady nie istnieje. Estymatę wartości prawdziwej opisuje niepewność wyniku pomiaru, która jest efektem niepewności wynikającej z błędów przypadkowych i z niedoskonałej korekcji błędów systematycznych [134]. Metoda wyznaczania niepewności pomiaru opiera się na określeniu rozkładów prawdopodobieństwa, a jej składowe mogą być opisywane za pomocą wariancji lub odchyleń standardowych.

W części eksperymentalnej rozprawy, w opisywanym systemie klasyfikacyjnym do celów pomiarowych zastał wykorzystany Kompaktowy Spektrometr CCS100, 350–700 nm, firmy Thorlabs. Wykonano 10 pomiarów testowych rejestrując widmo źródła światła o znanej charakterystyce spektralnej. Średniokwadratowa odchyłka od wartości średniej wynosiła +/- 0.01 rejestrowanej, unormowanej do jedności, bezwymiarowej wartości intensywności. Jest to jednoznaczne z odchyleniem +/- 1%. Uśredniona dla badanych długości fal niepewność pomiarowa wyrażona odchyleniem standardowym eksperymentalnym z 10 pomiarów wynosi 0.01. Wartość została obliczona na podstawie wzoru 3.4, jak podaje Główny Urząd Miar – GUM 2008, będącego najlepszą estymatą określającą liczbowo jak dobrze dana funkcja estymuje wartość oczekiwaną.

$$s(\lambda_{1\div k}) = \frac{\sum_{\lambda_1}^{\lambda_k} \sqrt{\frac{\sum_{j=1}^{n}(q_j - \overline{q})^2}{(n-1)}}}{k} \tag{3.4}$$

Gdzie:

$s(\lambda_{1\div k})$ – uśrednione dla badanych długości fal odchylenie standardowe eksperymentalne;

λ_1 – początkowa długość fali zarejestrowanego zakresu;

λ_k – końcowa długość fali zarejestrowanego zakresu;

q_j – pojedyncza obserwacja dla konkretnej długości fali;

$\overline{q}$ – średnia arytmetyczna n pomiarów dla konkretnej długości fali;

Tabela 3.1: Wyniki acc uzyskane w symulacji trzech zbiorów danych o odchyłkach od wartości średniej: +/- 2%, +/- 5% +/- 10% przy wykorzystaniu klasyfikacji metodą PAW. Wytłuszczono oryginalny pomiar: +/- 1%.

odchyłka od wartości średniej [%]	acc [%]
+/- 1	**89**
+/- 2	80
+/- 5	51
+/- 10	30

n – liczba powtórzonych pomiarów dla konkretnej długości fali;

k – liczba zarejestrowanych długości fal.

W celu zbadania wpływu niedokładności pomiaru na końcowy wynik klasyfikacji podczas korzystania z metody PAW, wykonano symulację sygnałów o średniokwadratowych odchyłkach od wartości średniej: +/- 2%, +/- 5% +/- 10%. Podstawą zamodelowanych widm były rzeczywiste pomiary próbek czterech typów miodów po filtracji za pomocą filtru S-G (szczegółowy opis eksperymentu znajduje się w podrozdziale 4.2.2). W opisywanej symulacji rozdzielczość widmowa pozostała niezmieniona w stosunku do oryginalnej. Otrzymane wyniki zaprezentowano w tabeli 3.1.

W miarę wzrostu odchyłki od średniej wartości rejestrowanej przez urządzenie dokładność klasyfikacji przy użyciu metody PAW maleje. Można przyjąć, że zadowalające efekty (acc powyżej 80%) uzyskuje się w pomiarach, których odchyłka nie przekracza +/- 2%.

W przypadku korzystania z technik spektralnych oprócz niedokładności pomiaru, w rozważaniach należy uwzględnić również rozdzielczość pomiarową w dziedzinie spektralnej urządzenia. Zgodnie z definicją podaną przez GUM 2009, rozdzielczością nazywa się najmniejszą możliwą statystycznie istotną różnicę pomiędzy bieżącym wskazaniem, a kolejnym (statystycznie nowym) wskazaniem konkretnego przyrządu. Wielkość tę można również rozumieć, jako wyznacznik precyzji pomiaru. Jest on nierozłącznie związany z procesem kwantyzacji wyniku, na który składają się poprawka o charakterze przypadkowym i poprawka o charakterze systematycznym. Rozdzielczość widmowa często określana jest poprzez szerokość połówkową np. filtrów, w przypadku kamer hierspektralnych, co ma bezpośredni wpływ na liczbę kanałów [8]. W części eksperymentalnej opisane jest użycie spektrometru o rozdzielczości pomiarowej w dziedzinie spektralnej 0.1 nm.

Tabela 3.2: Wpływ rozdzielczości pomiarowej w dziedzinie spektralnej zamodelowanych zbiorów danych na dokładności klasyfikacji (acc) uzyskane metodą PAW, dla różnych niepewności pomiarowych. Wytłuszczono oryginalny pomiar.

odchyłka od wartości średniej [%]	rozdzielczość pomiarowa [nm]	acc [%]
+/- 1	**0.1**	**89**
+/- 1	0.5	79
+/- 1	1	60
+/- 1	5	53
+/- 1	10	51
+/- 2	1	65
+/- 2	5	61
+/- 2	10	57
+/- 5	1	42
+/- 5	5	40
+/- 5	10	35
+/- 10	1	45
+/- 10	5	32
+/- 10	10	30

W celu wskazania zależności pomiędzy wartością rozdzielczości widmowej pomiaru, a wynikiem działania metody PAW, wykonano symulację 3 sygnałów o różnej liczbie kanałów spektralnych. Bazą modeli, jak w przypadku modeli odchyłki wartości od średniej, były rzeczywiste, przefiltrowane pomiary spektralne próbek miodów opisane w części 4.2.2. Sygnały zasymulowano poprzez wycięcie części danych oryginalnych tak, aby ich rozdzielczości pomiarowe w dziedzinie widma wynosiły odpowiednio: 1 nm, 5 nm i 10 nm (dla przypadku odchyłki +/- 1% dołączono symulację rozdzielczości 0.5 nm). Otrzymane wyniki zaprezentowano w tabeli 3.2.

Rozdzielczości pomiarowe o wartości większej niż 0.5 nm nie zapewniają wystarczająco wysokiej jakości danych, które umożliwiają klasyfikację z dokładnością powyżej 80% (podczas stosowania metody PAW). Przy odchyłce rzędu +/- 1% rozdzielczość pomiarowa w dziedzinie widma poniżej 0.5 nm zapewni względnie wysokie acc (powyżej 80%).

3.4. Porównanie metod selekcji cech

Tabela 3.3 przedstawia kompleksowe porównanie podstawowych algorytmów używanych do selekcji cech danych spektralnych - zarówno tych generujących nowe atrybuty sygnału (podrozdział 2.2.3.1), jak i tych wykorzystujących selekcję widmową (podrozdział 2.2.3.2)- w połączeniu z trzema klasyfikatorami: DT, będącą podstawową metodą klasyfikacyjną umożliwiającą interpretację wyników oraz powszechnie wykorzystywanymi w dziedzinie ANN i SVM. Analizie poddano następujące metody redukcji wymiarowości danych: LDA, PCA, PLS, SFFS, manualną selekcję usuwanych kanałów spektralnych oraz zaproponowane przez autorkę metody PAW i DRW. LDA może być wykorzystywana również, jako klasyfikator, dlatego w porównaniu występuje pojedynczo lub w połączeniu tylko z selekcją manualną, lub tylko z SFFS.

LDA, podobnie jak PCA, nie można stosować, gdy liczba cech jest większa od liczby próbek. Analogicznym obwarowaniem cechują się klasyfikatory ANN i SVM. Zagadnienie to nazywane jest w literaturze problemem małej ilości próbek [135]. Drzewa decyzyjne, jako proste algorytmy, nie wymagają do prawidłowego działania znacznej liczby próbek. Również PLS przystosowany jest do pracy na tzw. "fat data", czyli z danymi, w których liczba zmiennych znacznie przewyższa liczbę próbek.

W celu wykonania skutecznej selekcji cech za pomocą PCA, jak i PLS, dane muszą być ze sobą skorelowane. Powstałe nowe zmienne – składowe główne, w przypadku PCA – nie są już skorelowane względem siebie. Gdy nie występuje korelacja pomiędzy danymi wejściowymi, PCA i PLS nie zapewniają możliwości redukcji danych przy ograniczonej stracie informacji [76]. Spełnienie tego założenia nie jest wymagane podczas manualnej selekcji kanałów, czy przy użyciu SFFS.

Złożoność obliczeniowa konkretnych technik w przypadku dużej liczby zmiennych może mieć wpływ na płynność i czas wykonywanych kalkulacji. Najmniejszą złożonością obliczeniową charakteryzują się metody selekcji manualnej zakresów spektralnych. W takich przypadkach moc obliczeniowa zużywana jest jedynie na wykorzystanie modeli klasyfikacyjnych. Przykładem algorytmu o wysokiej złożoności obliczeniowej jest SFFS, który po każdorazowym dodaniu cechy do szukanego podzbioru, sprawdza, czy któraś z wcześniej wybranych cech nie pogarsza wartości funkcji kryterialnej. W pesymistycznym przypadku SFFS może dokonać wykładniczej ilości operacji maksymalizacji funkcji kryterialnej.

Interpretacja fizyczna nowo wygenerowanych cech redukujących dane wejściowe, może stwarzać problemy. Przykładowo wykorzystanie metody SFFS często skutkuje trudnościami w ocenie, w przypadku, gdy wynikiem będą bardzo wąskie zakresy spektralne. W przypadku PCA do interpretacji składowych głównych wielu autorów rekomenduje wykorzystanie współczynników korelacji między zmiennymi pierwotnymi, a daną składową główną, jednak nawet one nie dostarczają wielowymiarowej informacji odnośnie łącznego wkładu zmiennych wejściowych do danej składowej głównej [136]. Metoda PLS jest krytykowana głównie ze względu na trudny do interpretacji model. Również problematyczne są do interpretacji zależności pomiędzy predyktorami, a macierzą odpowiedzi [137]. Wykorzystanie podczas klasyfikacji ANN lub SVM umożliwia uzyskanie wyniku np. w postaci wartości dokładności klasyfikacji. Jednak nie ma on bezpośredniego przełożenia na informacje, które zakresy spektralne były znaczące podczas klasyfikacji, co jest istotne podczas próby adaptacji systemu do konkretnego zastosowania. Przy użyciu funkcji dyskryminacyjnych, takich jak LDA, obowiązuje zasada: im wyższa wartość standaryzowanego współczynnika, tym większy udział danej zmiennej w dyskryminowaniu grup – właściwość ta przekłada się na możliwości interpretacyjne modeli wykorzystujących LDA. Podczas paramteryzacji sygnału metodą PAW generowana jest wartość a mająca ścisłą definicję matematyczną odnoszącą się do wartości tangensa kąta nachylenia prostej dopasowanej do fragmentu wykresu. Metoda DRW pozwala na wycięcie konkretnych, kanałów spektralnych. Obie metody wykorzystują DT, które umożliwia wykonanie interpretacji fizycznej wyniku.

Stosowanie ANN, jak i SVM nie daje możliwości kontroli procesu decyzyjnego. Klasyfikatory te działają na zasadzie "czarnej skrzynki" w przeciwieństwie do DT. W przypadku algorytmów bazujących na DT operator ma wgląd w poszczególne poziomy klasyfikacji i dzięki temu większą kontrolę nad całym procesem.

Uznaje się, że selekcja cech za pomocą LDA, PCA, PLS i SFFS jest na tyle automatyczna, że operator faktycznie nie oddziałuje na wynik i jego doświadczenie nie jest wymagane. Podczas selekcji kanałów spektralnych w sposób manualny doświadczenie operatora jest niezbędne. Również techniki klasyfikacyjne wymagają wcześniejszej praktyki, choć technika DT, z tych wymienionych w tabeli, wydaje się być najbardziej intuicyjna. W przypadku korzystania z sieci neuronowych użytkownik ma możliwość kontroli procesu poprzez wybór typu sieci, określenie liczby warstw i neuronów lub implementacji gotowego modelu. Podczas użycia

SVM należy wskazać typ jądra i określić jego parametry lub zastosować istniejące rozwiązanie.

Wiele przedstawionych w niniejszej pracy technik bazuje na zaawansowanej matematyce i wykorzystuje najnowsze technologie z dziedziny IT. Użycie skomplikowanych algorytmów, może wiązać się z faktem, iż użytkownik stosuje modele jako "czarne skrzynki" przestając kontrolować zachodzące w nich procesy. W konsekwencji uzyskany wynik jest trudny do interpretacji fizycznej, co jest równoznaczne ze znacznym ograniczeniem możliwości wykorzystania go w celu adaptacji systemu pomiarowego do konkretnego zastosowania. Dostosowując zakresy pracy urządzeń, dobierając konkretne, bardziej dopasowane do zadania elementy mechaniczno - optyczne zmniejsza się koszty produkcji i pracy systemu.

Metody PAW oraz DRW w połączeniu z klasyfikatorem w postaci DT mogą znaleźć zastosowanie w prototypowaniu inżynierskich systemów do klasyfikacji ze względu na swoją transparentność i klarowność przeprowadzanych procesów. Możliwość ich wykorzystania w przypadku większej liczby cech niż liczby próbek, mimo prostoty zaimplementowanych algorytmów, daje szerokie spektrum zastosowań obu metod.

3.5. Podsumowanie rozdziału

Przedstawione w niniejszym rozdziale metody selekcji cech PAW i DRW są podstawą działania części algorytmicznej prezentowanego w rozprawie systemu do klasyfikacji obiektów warstwowych wykorzystującego techniki spektralne VIS. W celu rozpoczęcia procedury wykonania powyższych metod konieczna jest parametryzacja przygotowanych wcześniej danych widmowych. Na początku następuje ich akwizycja za pomocą urządzenia rejestrującego spektrum fal elektromagnetycznych, następnie sygnał próbki zostaje uniezależniony od innych elementów układu (obliczenie transmitancji lub reflektancji sygnału), po czym wykonywana jest filtracja wysokich częstotliwości. Po zakończeniu tych czynności dane mogą zastać poddane algorytmom metody PAW. Następuje wygenerowanie współczynników kierunkowych prostych metodą najmniejszych kwadratów dla wielomianów stopnia co najwyżej pierwszego, dopasowanych do odpowiednio dobranych zakresów spektralnych wykresu transmitancji (lub reflektancji). Tak powstaje dwuwymiarowa macierz wartości a z przyporządkowaną każdemu elementowi odpowiadającą mu długością fali.

Wykorzystanie serii współczynników kierunkowych prostych do opisu charakterystyki widmowej próbki jest zaletą przedstawianego rozwiązania. Taki sposób zapisu umożliwia

Tabela 3.3: Porównanie wybranych metod selekcji cech z uwzględnieniem redukcji wymiarowości poprzez generację nowych cech i selekcję widmową w połączeniu z wybranymi klasyfikatorami. LDA – liniowa analiza dyskryminacyjna, PCA – analiza składowych głównych, PLS – metoda cząstkowych najmniejszych kwadratów, SFFS – algorytm ruchomej selekcji postępującej, DT – drzewo decyzyjne, ANN – sztuczne sieci neuronowe, SVM – maszyna wektorów nośnych.

selekcja cech		klasyfikator	możliwość wykonania gdy liczba cech ≥liczba próbek	dane nie muszą być ze sobą skorelowane	złożoność obliczeniowa 1–3 (mała–duża)	możliwość interpretacji fizycznej wyników	możliwość kontroli procesu decyzyjnego na każdym etapie	niewymagane doświadczenie operatora
redukcja wymiarowości — generacja nowych cech	redukcja wymiarowości — selekcja widmowa							
LDA			✗	✓	2	✓	w małym zakresie	✓
LDA	selekcja manualna		✗	✓	2	✓	w małym zakresie	✗
LDA	SFFS		✗	✓	3	✓	w małym zakresie	✓
PCA		+ DT	✗	✗	2	✗	w małym zakresie	✗/✓
PCA		+ ANN	✗	✗	3	✗	✗	✗
PCA		+ SVM	✗	✗	2	✗	✗	✗
PCA	SFFS	+ DT	✗	✗	3	✗	w małym zakresie	✗/✓
PCA	SFFS	+ ANN	✗	✗	3	✗	✗	✗
PCA	SFFS	+ SVM	✗	✗	3	✗	✗	✗
PLS		+ DT	✓	✗	2	✗	w małym zakresie	✗/✓
PLS		+ ANN	✗	✗	3	✗	✗	✗
PLS		+ SVM	✗	✗	2	✗	✗	✗
PLS	SFFS	+ DT	✓	✗	3	✗	w małym zakresie	✗/✓
PLS	SFFS	+ ANN	✗	✗	3	✗	✗	✗
PLS	SFFS	+ SVM	✗	✗	3	✗	✗	✗
	selekcja manualna	+ DT	✓	✓	1	✓	✓	✗
	selekcja manualna	+ ANN	✗	✓	3	✗	✗	✗
	selekcja manualna	+ SVM	✗	✓	2	✗	✗	✗
	SFFS	+ DT	✓	✓	3	✓	✓	✗/✓
	SFFS	+ ANN	✗	✓	3	✗	✗	✗
	SFFS	+ SVM	✗	✓	3	✗	✗	✗
metoda PAW	metoda DRW	+ DT	✓	✗	2	✓	✓	✗/✓
metoda PAW	metoda DRW	+ ANN	✗	✗	3	✗	✗	✗
metoda PAW	metoda DRW	+ SVM	✗	✗	2	✗	✗	✗

uniezależnienie danych od zarejestrowanych konkretnych wartości intensywności, które stają się problematyczne podczas pomiarów zmiennej grubości próbek. Przykładowo pomiar skorup o analogicznych właściwościach morfologicznych, lecz różnych grubościach, umożliwi rejestrację widm zbliżonych do siebie kształtem, lecz rozsuniętych na osi pionowej. Efekt ten można w sposób manualny zminimalizować poprzez dostosowanie urządzenia pomiarowego do konkretnej próbki (w przypadku kompaktowego spektrometru siatkowego jest to tzw. czas integracji). Podczas procesu klasyfikacji ważne jest, aby zróżnicowanie wewnątrzgrupowe było jak najmniejsze. Spełnienie tego warunku może okazać się problematyczne w sytuacji uwzględnienia rozsunięcia po osi y podobnych kształtem wykresów transmitancji. Z tego względu parametryzacja sygnałów za pomocą wielkości określających jedynie nachylenie poszczególnych fragmentów widma pozwala użytkownikowi na pewną dowolność w kwestii grubości próbki.

Metoda najlepiej sprawdza się na sygnałach o niskich poziomach niepewności (uśrednione dla badanych długości fal odchylenie standardowe eksperymentalne rzędu 0.01) i rozdzielczości pomiarowej o wartości poniżej 0.5 nm.

Zakończenie redukcji wymiarowości danych po wykonaniu metody PAW może być wystarczające do uzyskania satysfakcjonującego rezultatu klasyfikacji. Jednak w przypadku, gdy metoda PAW nie przyniesie odpowiednio wysokiego poziomu przyporządkowania w procesie klasyfikacji, wskazane jest wykonanie kolejnej procedury – metody DRW. Jej działanie polega na usunięciu danych, w oparciu o algorytm ruchomej selekcji postępującej (SFFS). Procedura SFFS została zaadoptowana do zastosowań fizycznych poprzez wykorzystanie okien wycinających o szerokich pasmach spektralnych. Dzięki tak zaprojektowanej metodologi redukcji wymiarowości danych możliwe jest dostosowanie układu optycznego do konkretnego zadania, za pomocą optymalizacji liczby oraz zakresu zbieranych danych za pomocą fizycznych elementów układu takich jak filtry czy odpowiednio dobrane diody elektroluminescencyjne o wyselekcjonowanych zakresach spektralnych.

Obie przedstawione metody redukcji wykorzystują w swoich algorytmach klasyfikatory DT dokonujące ostatecznego procesu przyporządkowania obiektów do poszczególnych klas.

Niniejszy rozdział zawiera również kompleksowe podsumowanie popularnych metod selekcji cech w kontekście proponowanych w rozprawie algorytmów.

4. Adaptacyjność systemu wykorzystującego metody PAW i DRW na przykładach

4.1. System

Prezentowany w rozprawie system do klasyfikacji obiektów warstwowych składa się z następujących części (rysunek 4.1):

- konstrukcyjnej – układ optyczny przeznaczony do rejestracji sygnałów spektralnych VIS pochodzących z badanego materiału;
- obliczeniowej – przetwarzanie wstępne, wykorzystanie metody PAW lub metod PAW i DRW;
- klasyfikacyjnej – klasyfikacja obiektów do klas za pomocą drzewa decyzyjnego.

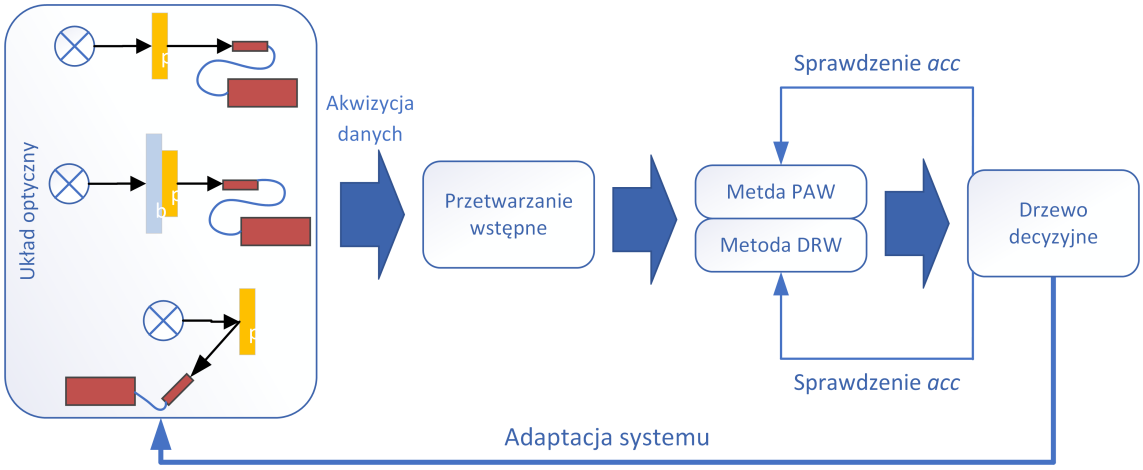
Idea stojąca u podstaw działania proponowanej metody i wykorzystane elementy pozwalają na łatwość modyfikacji systemu. Oznacza to, że nie ograniczają go niestandardowe kształty, czy rozmiary próbek. Dowodzą tego zaprezentowane w niniejszym rozdziale przykłady wykorzystania systemu. Możliwe ustawienia elementów konstrukcyjnych przedstawiono w części 2.2.2.1.

W pracy przedstawiono przypadek, w którym posłużono się obiema metodami – PAW i DRW, jak również opisano badanie wskazujące, że zastosowanie jedynie metody PAW może być wystarczające. Szczegółowy opis metod znajduje się w podrozdziałach 3.2.1 i 3.2.2. Wykorzystanie którejkolwiek z wymienionych metod implikuje zastosowanie DT, które jest jednocześnie klasyfikatorem niniejszego systemu.

4.2. Przykłady zastosowań systemu

Opis eksperymentów został podzielony na następujące sekcje:

- wstęp;



Rysunek 4.1: Schemat idei działania opracowanego systemu do klasyfikacji obiektów warstwowych wykorzystującego techniki spektralne VIS.

- próbki;
- układ optyczny;
- analiza i wyniki działania metody PAW – przypadek eksperymentu 1., metody PAW i DRW – w przypadku eksperymentu 2.;
- porównanie wyników działania innych metod;
- wnioski.

Pierwszy prezentowany eksperyment dotyczy klasyfikacji skorup jaj kurzych ze względu na efekt działania patogenu Mycoplasma Synoviae na zainfekowane zwierzę.

Drugi eksperyment prezentuje klasyfikację miodów ze względu na pochodzenie botaniczne. Miód jest materiałem wymagającym naniesienia na bazę. Jego konsystencja wymogła przekonstruowanie optyki systemu na układ działający w pionie.

Wykorzystanie systemu używającego metodę PAW zarówno w pierwszym jak i w drugim eksperymencie, implikowało wykonanie następujących czynności:

1. obliczenie widmowej transmitancji badanego materiału, przy jednoczesnej eliminacji wpływu charakterystyki podłoża i źródła światła na badaną próbkę - podrozdział 2.2.2.1;
   - w przypadku badań miodu konieczne jest wykonanie dodatkowych pomiarów intensywności zarejestrowanej po przejściu przez niepokrytą badaną substancją bazę;
2. zaimplementowanie metody PAW z zastosowaniem klasyfikatora DT z 7-krotną kroswalidacją i obliczenie acc oraz wykonanie walidacji prostej i obliczenie $acc\ r.$ – podrozdział 3.2.1 i 2.2.5;

3. w przypadku badań miodu, zaimplementowanie metody DRW z zastosowaniem klasyfikatora DT z 7-krotną kroswalidacją i obliczenie acc oraz wykonanie walidacji prostej i obliczenie $acc\ r.$ – podrozdział 3.2.2 i 2.2.5;
4. adaptacja systemu do konkretnego zastosowania:
   - w eksperymencie 1. – klasyfikacja skorup jaj kurzych na $chore$ i $zdrowe$;
   - w eksperymencie 2. – klasyfikacja miodów pod względem pochodzenia botanicznego.

Wybór technik porównywanych

Wyniki uzyskane za pomocą zaproponowanych przez autorkę metod zostały porównane z algorytmami powszechnie wykorzystywanymi poprzez przeprowadzenie następujących obliczeń:

- redukcja wymiarowości danych spektralnych za pomocą PCA i obliczenie acc z wykorzystaniem klasyfikatora DT z 7-krotną kroswalidacją oraz wykonanie walidacji prostej i obliczenie $acc\ r.$ – podrozdział 2.2.3.1;
- RBF na 3 i 10 składowych głównych PCA oraz wykonanie walidacji prostej i obliczenie $acc\ r.$ – podrozdział 2.2.4;
- SVM na 3 i 10 składowych głównych PCA oraz wykonanie walidacji prostej i obliczenie $acc\ r.$ – podrozdział 2.2.4;
- SVM na danych będących transmitancją oraz wykonanie walidacji prostej i obliczenie $acc\ r.$

Metoda PCA jako jedna z najczęściej wykorzystywanych metod redukcji wymiarowości wydaje się być oczywistym wyborem do celów porównawczych. Połączenie jej z klasyfikatorem DT daje szansę bezpośredniego porównania z efektem obliczeń metody PAW.

Wybór sieci RBF został podyktowany jej powszechnym wykorzystaniem w zagadnieniach klasyfikacji obiektów. Przykładami zastosowań tej sieci jest analiza danych pochodzących np. ze spektroskopii VIS-NIR wykonująca klasyfikację upraw i chwastów [141], klasyfikacji odmian herbat – zielona, czarna i ulong (przy wykorzystaniu spektroskopii NIR z równoczesnym użyciem maszyny wektorów nośnych) [142] lub klasyfikacji obrazów multispektralnych [143].

Ze względu na brak konieczności ograniczenia do separowalności liniowej, kolejnym algorytmem, który został przetestowany jest SVM. SVM umożliwia przeprowadzenie obliczeń bezpośrednio na danych transmitancji, co zostało również zaimplementowane.

Ponadto, ze względu na charakter badania, dodatkowo zostały wykonane następujące zadania (w opisie eksperymentu oznaczone symbolem *):

- w eksperymencie związanym z badaniem skorup jaj – selekcja manualna kanałów spektralnych i parametrów a na podstawie wykresu transmitancji w zależności od długości fali wraz z wykonaniem DT z 7-krotną kroswalidacją i obliczenie acc – podrozdział 2.2.4 i 2.2.5;
- w eksperymencie związanym z badaniem miodów – redukcja wymiarowości wygenerowanych parametrów a za pomocą PCA i obliczenie acc przy użyciu klasyfikatora DT z 7-krotną kroswalidacją oraz wykonanie walidacji prostej i obliczenie $acc\ r$.

Wykresy transmitancji skorup jaj kurzych z eksperymentu 1. są na tyle specyficzne, że możliwe jest przeprowadzenie ręcznej selekcji kanałów spektralnych i dalsza ich analiza. Ze względu na niemożność manualnej selekcji kanałów w eksperymencie z miodami, wykonano analizę PCA bazującą na parametrach a obliczonych w trakcie działania metody PAW. W celu łatwości porównania wykonano klasyfikację DT wyników obliczając $acc\ r$.

Wartości acc są uśrednione z 50 powtórzeń. Algorytmy powszechnie używane zostały zaimplementowane w $C\sharp$ z wykorzystaniem biblioteki Encog (RBF i SVM) i Accord (PCA).

4.2.1. Klasyfikacja skorup jaj kurzych ze względu na efekt działania Mycoplasma Synoviae

Wstęp

Kury nioski na terenie Polski objęte są kontrolą weterynaryjną podlegającą Inspekcji Weterynaryjnej na danym terenie. Działania typu oględziny i interwencje oparte są m.in. o Rozporządzenie Ministra Rolnictwa i Rozwoju Wsi z dn. 8 lutego 2019 r. i wcześniejszych, w sprawie wprowadzenia „Krajowego programu zwalczania niektórych serotypów Salmonella w stadach kur niosek gatunku Gallus gallus" na lata 2019 i 2020 (Dz. U. 2019, poz. 346). Finlandia i Szwecja już w 2004 roku w odpowiedzi na epidemię ptasiej grypy (szczep H5N1, okres epidemii 2003–2006) poszerzyły zakres badanych zakażeń nieograniczając się jedynie do typu Salmonella.

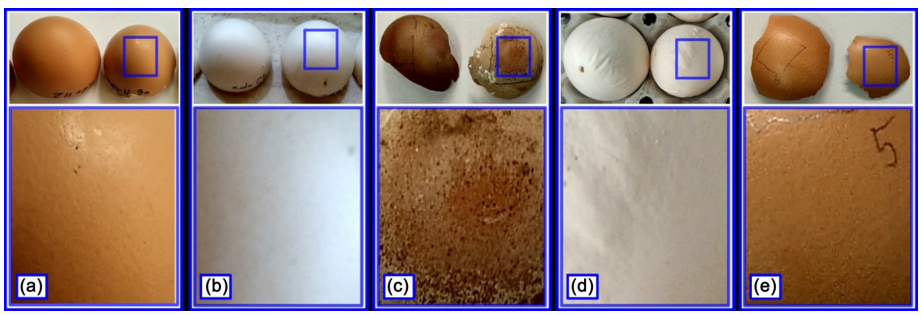
Mycoplasma Synoviae (MS) jest jednym z głównych patogenów występujących w stadach drobiu hodowlanego powodującym subkliniczne zakażenia. MS może powodować poważne infekcje układu oddechowego, problemy z prawidłowym funkcjonowaniem zatok u zwierząt, jak również prowadzi do „nieprawidłowości budowy strukturalnej wierzchołka skorupy jaja". Zdeformowana, niejednorodna skorupa jest podatna na pęknięcia, przez które dostają się drobnoustroje. Budowa i skład jaja blokują bakteriom dostęp do wnętrza. Z zewnątrz jajo pokryte jest woskową kutikulą utrudniającą dostęp mikroorganizmów, a skorupa i błony stanowią naturalne filtry blokujące wejście bakterii. W białku i żółtku można znaleźć wiele substancji (albuminę, lizozym, cystatynę, immunoglobulinę) działających bezpośrednio lub pośrednio na patogeny i hamujących procesy zapalne. Mimo tak dobrej ochrony przypadki zatrucia skażonymi jajami nie należą do rzadkości. Dotyczy to zwłaszcza szkodliwych dla zdrowia ludzkiego szczepów Salmonella. Przypuszcza się, że przyczyną jest uszkodzenie lub nieprawidłowa budowa warstw ochronnych jaja, ze szczególnym uwzględnieniem nieszczelności skorupy, której struktura może być naruszona z różnych przyczyn, a jedną z nich niewątpliwie jest działanie MS.

Znaczna część skorup jaj pochodzących od kur zakażonych MS nie cechuje się charakterystycznym odkształceniem skorupy i dlatego niektóre z nich mogą zostać nieprawidłowo sklasyfikowane podczas rutynowej kontroli wykonywanej przez człowieka. Zaproponowane w rozprawie rozwiązanie pozwala zminimalizować powyższe ryzyko. Przedstawiony system wykorzystuje technikę spektralną VIS do klasyfikacji skorup jaj kurzych, rozumianych jako

obiekty warstwowe. Około 0.1% strumienia światła docierającego do skorupy przechodzi przez nią. Pozostała część jest rozpraszana lub pochłaniana przez cząsteczki badanego materiału. Jednakże zarejestrowany sygnał, stanowiący podstawę do obliczenia transmitancji próbek jest wystarczający, aby wskazać znaczące różnice charakterystyk pomiędzy skorupami $chorymi$ i $zdrowymi$. Opisany proces daje możliwość prawidłowego przyporządkowania badanej próbki do grupy skorup o podwyższonym ryzyku z dokładnością na poziomie 96%. W przypadku zarejestrowania kilku pozytywnych wyników testu na występowanie podejrzenia zakażenia kury patogenem MS stado kwalifikuje się do bardziej szczegółowych ukierunkowanych badań. Regularne wykonywanie tego typu testów może wspomóc wczesne wykrycie występowania bakterii MS w stadach drobiu. To z kolei zmniejsza koszty związane z antybiotykoterapią lub ewentualnym ubojem sanitarnym, wyczyszczeniem i zdezynfekowaniem zakładu.

Materiał badawczy

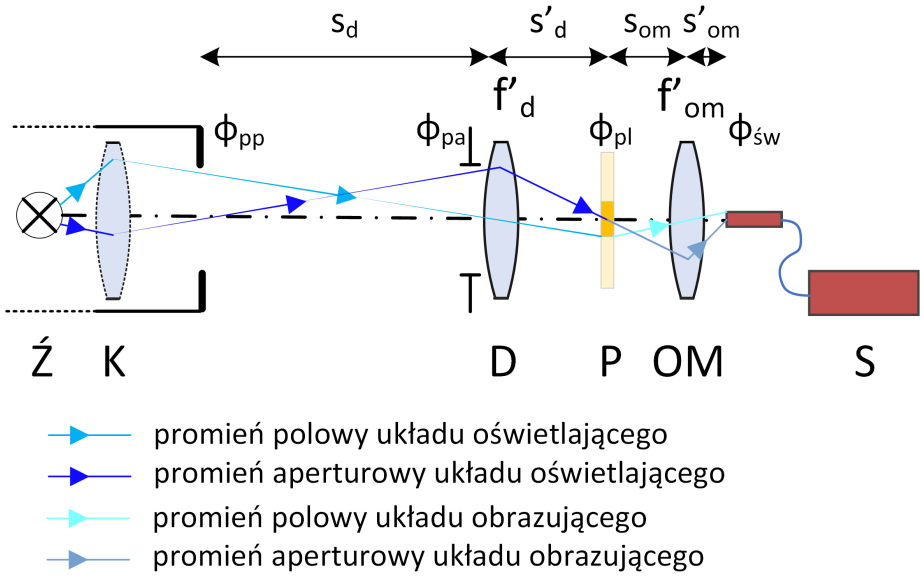
Testom zostało poddane 60 brązowych i 60 białych skorup jaj kurzych. W każdej grupie połowa jaj pochodziła od kur zakażonych SM, a połowa od kur zdrowych. Wszystkie próbki zostały dostarczone przez Państwowy Instytut Weterynarii w Puławach. Zainfekowane ptaki miały objawy związane z infekcjami mykoplazmatycznymi, lecz nie wszystkie skorupy pochodzące od chorych jednostek – w dalszej części tekstu skrótowo nazywane skorupami $chorymi$ – charakteryzowały się "nieprawidłowościami wierzchołka skorupy jaja". Skorupy pochodzące od zdrowych zwierząt zostały dostarczone z tzw. grup kontrolnych – w dalszej części tekstu skrótowo nazywane skorupami $zdrowymi$. Przykłady badanych skorup przedstawione są na rysunku 4.2. Przygotowanie próbki do eksperymentu polegało na: rozbiciu jaja, przepłukaniu skorupy pod strumieniem letniej, bieżącej wody i pozostawieniu do wyschnięcia na 3 dni.



Rysunek 4.2: Przykłady analizowanych próbek skorup jaj kurzych: (a) brązowe $zdrowe$ skorupy; (b) białe $zdrowe$ skorupy; (c) brązowe $chore$ skorupy; (d) białe $chore$ skorupy; (e) $chora$ skorupa bez widocznych deformacji [140].

Układ optyczny

Zbudowany układ optyczny bazuje na idei pomiaru transmitancji próbki przedstawionej na rysunku 2.2 (b) w rozdziale 2.2.2.1 (układ niewykorzystujący bazy). Szczegółowy schemat wykonanego układu optycznego zaprezentowano na rysunku 4.3. Schemat nie został wykonany w skali. Można wyróżnić część oświetlającą składającą się z elementów na lewo od próbki oraz część obrazującą przedstawioną po prawej stronie badanego obiektu. Skorupa jaja kurzego ma charakter rozpraszający światło, co determinuje rozróżnienie biegu promieni polowego i aperturowego osobno w części oświetlającej i obrazującej (na rysunku przedstawiono to za pomocą różnych kolorów promieni).



Rysunek 4.3: Schemat układu optycznego do pomiarów transmitancji skorup jaj, bez zachowanej skali. Ź – stabilizowane inkadescencyjne źródło światła, K – kolektor, Φ_{pp} – średnica otworu przysłony polowej, D – dublet achromatyczny, jego oprawa pełni rolę źrenicy wejściowej będąc jednocześnie przysłoną aperturową (Φ_{pa}), f'$_d$ – ogniskowa dubletu, P – próbka (skorupa jaja kurzego) z zaznaczonym na pomarańczowo obszarem pomiarowym o średnicy Φ_{pl} plamki światła, OM – obiektyw mikroskopowy, f'$_{om}$ – ogniskowa soczewki symbolizującej obiektyw mikroskopowy, S – kompaktowy spektrometr z wejściem światłowodowym, Φ_w – średnica wejściowa światłowodu, s$_d$ – odległość dubletu od przysłony pola, s'$_d$ – odległość próbki od dubletu, s$_{om}$ – odległość próbki od obiektywu, s'$_{om}$ – odległość czoła światłowodu od obiektywu.

Część układu pełniąca rolę oświetlającą składa się z oświetlacza i dubletu achromatycznego (D). Inkandescencyjne źródło światła (Ź) oraz kolektor (K) tworzą oświetlacz. Wykorzystana lampa charakteryzuje się widmem ciągłym, w zakresie od 400 nm do końca zakresu pracy spektrometru – 750 nm. Jej maksymalna wartość intensywności odpowiada długości fali

640 nm. Wiązka promieni wydobywająca się z oświetlacza przechodzi przez otwór zwany przysłoną polową (PP) o średnicy Φ_{pp} 3 mm ograniczający rozmiary kątowe obrazowanego źródła światła. Dublet achromatyczny o ogniskowej (f'$_d$) 35 mm pracuje w powiększeniu -0.3. Jego oprawa pełni rolę źrenicy wejściowej, jednocześnie będąc przysłoną aperturową (Φ_{pa}). Układ oświetlający tworzy jednorodną plamkę świetlną o średnicy ok. 1 mm (Φ_{pl}) w płaszczyźnie badanej próbki (P).

Rejestrowany pomiar zawiera informacje będące efektem całkowania widma przepuszczonego przez tę powierzchnię skorupy. Kolejne elementy przez które przechodzi wiązka światła wchodzą w skład układu obrazującego. Promienie przepuszczone przez próbkę docierają do obiektywu mikroskopowego (OM) pracującego w nietypowym dla siebie układzie – obraz zostaje pomniejszony, a nie powiększony, jak w standardowym zastosowaniu mikroskopowym.

W celu wykonania prawidłowej konstrukcji układu optycznego wykonano obliczenia gabarytowe na przybliżeniu cienkosoczewkowym. Obiektyw mikroskopowy uznano za cienką soczewkę o ogniskowej (f'$_{om}$) 4 mm. Nominalna wartość powiększenia użytego obiektywu wynosi 50 x, jednak w niniejszym układzie pracuje on w powiększeniu -0.2, co oznacza, że obraz plamki światła pomniejszony jest pięciokrotnie wypełniając w całości powierzchnię czoła światłowodu o średnicy (Φ_w) 200 μm. Efekty wynikające z nieskompensowanych aberracji, które mogłyby się pojawić przy nietypowym ustawieniu obiektywu uznaje się za stałe dla całego pomiaru i pomijalne. Ich analiza na tym etapie badań nie jest przedmiotem rozprawy.

Obiektyw dobrano doświadczalnie, aby dopasować się energetycznie do jego apertury numerycznej (NA_{ob} = 0.55) i średnicy światłowodu. Wiązka światła jest doprowadzana przez światłowód do kompaktowego spektrometru Thorlabs CCS100 (S) działającego w zakresie VIS (350 nm–750 nm), rejestrującego strumień z rozdzielczością 0.1 nm. Analiza niepewności pomiarowych wykorzystanego urządzenia wraz z symulacjami znajduje się w podrozdziale 3.3.

W celu uzyskania optymalnego rozłożenia elementów optycznych i maksymalizacji rejestrowanego natężenia wiązki światła wykonano stosowne obliczenia uzyskując, a następnie implementując w rzeczywistym układzie następujące wartości:

- s_d - odległość dubletu od przysłony pola: -157 mm;
- s'_d - odległość próbki od dubletu: 45 mm;
- s_{om} - odległość próbki od obiektywu: -24 mm;
- s'_{om} - odległość czoła światłowodu od obiektywu: 4.8 mm.

Pomiary

Detekor w postaci kompaktowego spektrometru siatkowego VIS rejestruje intensywność wiązki światła po przejściu przez skorupę. Wykonano serie pomiarów punktowych z obszaru górnego i dolnego wierzchołka skorupy jaja oraz części bocznej. W przypadku widocznej deformacji skorupy (np. pofałdowanie) zmieniony fragment był traktowany jako dodatkowa powierzchnia do badań. Dla każdego obszaru pomiarowego zostało wykonane pięć rejestracji widma. Sumarycznie dokonano 2541 pomiarów spektralnych. Czas integracji jednorazowej akwizycji widma wynosił od 300 to 1000 ms i zależał od:

- barwy skorupy;
- grubości skorupy;
- kształtu próbek (niemożliwe ich powtórne umieszczenie w tej samej pozycji w układzie pomiarowym).

Analiza i wyniki

Niniejszy podrozdział został podzielony na sekcję dotyczącą analizy i wyników realizacji wykorzystania systemu z użyciem metody PAW oraz sekcję przedstawiającą rezultaty wykonania obliczeń porównawczych przy zastosowaniu innych algorytmów. Kalkulacje przeprowadzono w dwóch osobnych grupach – skorup brązowych i skorup białych. Podrozdział podsumowuje tabela zawierająca zebrane wyniki działań poszczególnych algorytmów.

Analiza realizacji wykorzystania systemu z użyciem metody PAW

Zarejestrowane sygnały intensywnościowe zostały poddane następującej procedurze:

1. obliczenie transmitancji materiału;
2. wykorzystanie metody PAW;

- obliczenie dokładności acc na podstawie DT z kroswalidacją

  Metoda PAW pozwoliła na automatyczne przetestowanie znacznej liczby przypadków dopasowania prostych do wykresów transmitancji przy jednoczesnym sprawdzaniu wyników acc wygenerowanych drzew;

Tabela 4.1: Podsumowanie liczby próbek skorup $chorych$ i $zdrowych$ będących składowymi zbioru uczącego i testowego w grupie skorup białych i brązowych.

		l. skorup chorych	l. skorup zdrowych
skorupy białe	zbiór uczący	615	376
	zbiór testowy	300	200
skorupy brązowe	zbiór uczący	375	340
	zbiór testowy	200	150

- sprawdzenie – obliczenie współczynnika dokładności $acc\ r.$ za pomocą walidacji prostej
  Dane będące efektem działania algorytmu paramateryzacji (wartości a), podzielono losowo na zbiór uczący i zbiór testowy, w proporcjach przedstawionych w tabeli 4.1.

Wyniki działania systemu z użyciem PAW

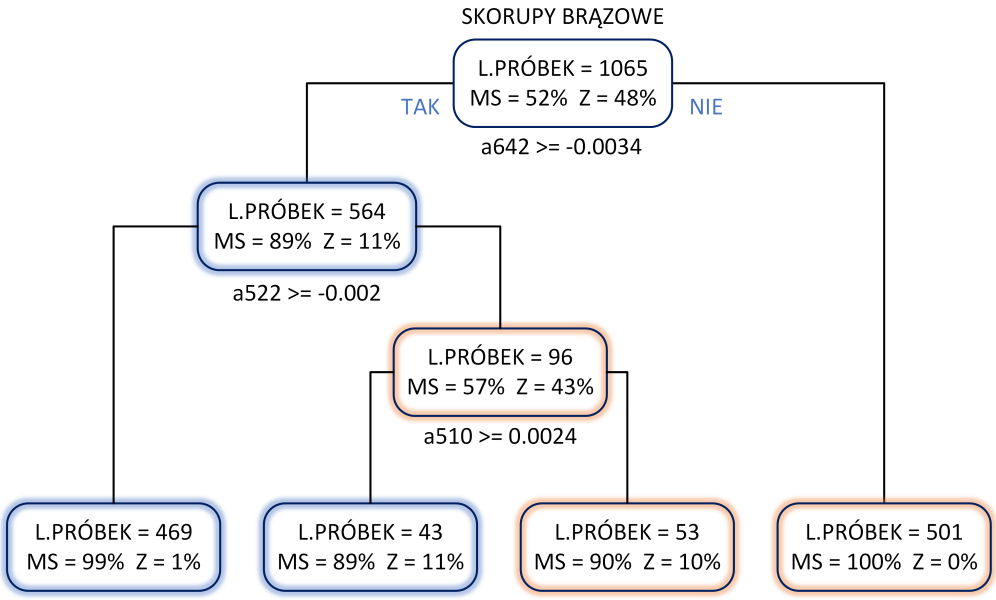
Wykorzystanie metody PAW pozwoliło na osiągnięcie wartości acc = 97% po 7-krotnej kroswalidacji na obu zbiorach próbek podczas zaimplementowania parametrów: $zakres$ = $skok$ = 5 nm oraz $zakres$ = $skok$ = 1 nm, odpowiednio w grupie skorup białych i brązowych. Modele generujące najwyższe dokładności dopasowania przedstawione są na rysunkach 4.4 i 4.5.

Podczas procesu sprawdzania uzyskanych wyników za pomocą walidacji prostej stosunek poprawnie sklasyfikowanych próbek do liczby wszystkich próbek ze zbioru testowego (parametr $acc\ r.$) w przypadku skorup białych wynosił 96%, a dla skorup brązowych 95%.

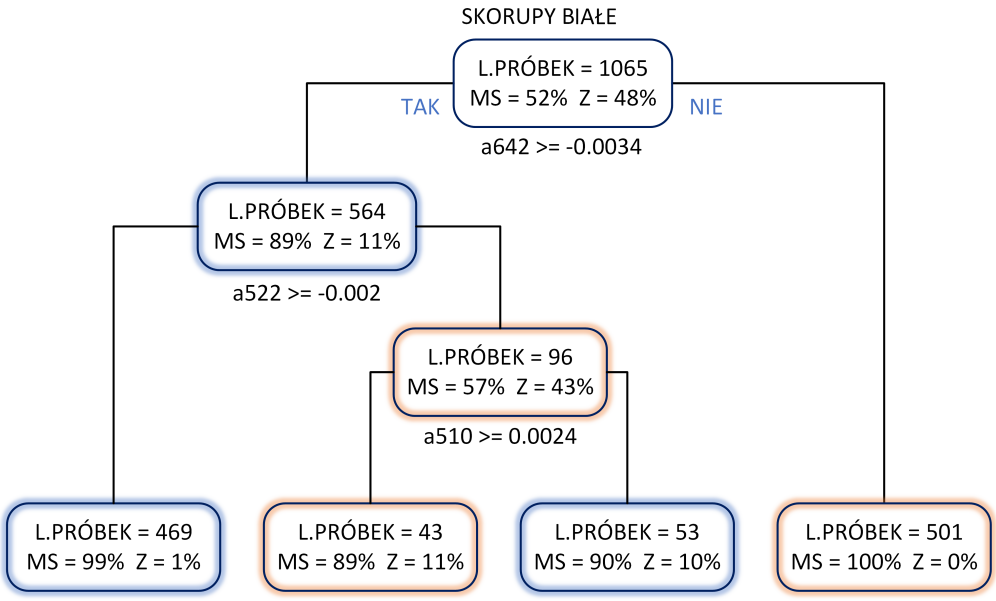
Analiza i wyniki działania innych algorytmów

1. PCA z danych spektralnych, obliczenie acc i $acc\ r.$

   Wykonano zmniejszenie wymiarowości danych za pomocą PCA. Na nowo powstałych wielkościach będących pierwszymi 3 składowymi głównymi PCA, wygenerowano drzewo decyzyjne. Wynikiem końcowym acc, po 7-krotnej kroswalidacji było osiągnięcie poziomu 96% i 88% odpowiednio dla dla grupy skorup białych i brązowych. W celu sprawdzenia uzyskanych wyników przeprowadzono walidację prostą na zbiorach takich, jak te przedstawione w tabeli 4.1. Dla obu grup uzyskano $acc\ r.$ równe 95%.



Rysunek 4.4: Wynik działania metody PAW w postaci drzewa decyzyjnego dla grupy skorup brązowych. $acc = 97\%$ po 7-krotnej kroswalidacji przy automatycznie dobranych parametrach: $zakres = skok = 5$ nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*.



Rysunek 4.5: Wynik działania metody PAW w postaci drzewa decyzyjnego dla grupy skorup białych. $acc = 97\%$ po 7-krotnej kroswalidacji przy automatycznie dobranych parametrach: $zakres = skok = 1$ nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*.

2. RBF na 3 i 10 składowych głównych PCA

   Obliczone składowe główne PCA posłużyły za dane wejściowe do zbudowania sieci RBF. Warstwa ukryta składała się z neuronów implementujących bazowe funkcje radialne będącymi w tym przypadku funkcjami Gaussa (wzór 4.1). Pojedyncza funkcja radialna, nazywana jądrem, charakteryzuje się parametrem σ – szerokość jądra, który wynosił 1 w powyższej implementacji. Odległości pomiędzy neuronami zawierają się w zakresie od 0 do 1. Sieć uczy się wykorzystując algorytm SVD (singular value decomposition).

$$G(r) = exp\left(\frac{-r^2}{2\sigma^2}\right) \tag{4.1}$$

   gdzie $r(x,c) = \|x - c\| = \sqrt{(x-c)^T(x-c)}$

   x – argument funkcji,

   c – położenie centrum funkcji.

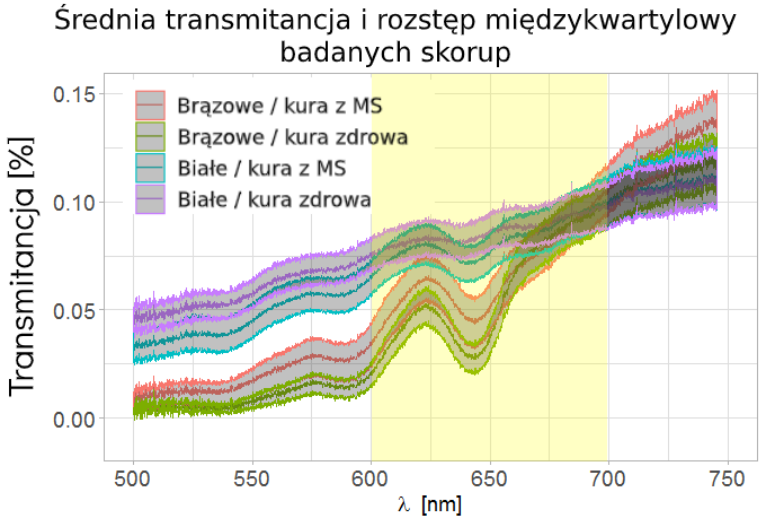
   Warstwa wyjściowa sumuje aktywacje neuronów warstwy ukrytej podając wartość sumy, jako końcowy wynik działania sieci.

   Wykonano dwie sieci RBF, w jednym przypadku wykorzystano 3 PC do zbudowania sieci, w drugim użyto 10 PC. Pierwszy przypadek charakteryzował się wyższymi wartościami $acc\ r.$ (95% i 92%, białe i brązowe skorupy) w porównaniu do drugiego (94% i 85%, białe i brązowe skorupy). Prawdopodobnie jest to uzasadnione faktem, iż wektory główne o wyższych indeksach opisują coraz więcej szumu. Oznacza to, że trzy pierwsze PC wystarczą do wydajnej klasyfikacji, a wykorzystanie kolejnych wprowadza dane obniżające dokładność $acc\ r.$

3. SVM na 3 składowych głównych PCA oraz na danych będących transmitancją

   Podczas implementacji SVM wykorzystano algorytm – klasyfikacja wektorów nośnych. Użyto gausowską funkcję jądra o parametrze γ równym $1/z$, gdzie z oznacza ilość wymiarów wejścia (różna w zależności od analizowanego przypadku). Funkcja stopu ustawiona jest na wartość $1e-3$. Po wykonaniu kilku różnych konfiguracji, wybrana wartość generalizującego parametru C wynosi 1. Uzyskano 96% dokładności podczas walidacji prostej 3 pierwszych składowych głównych PCA zarówno dla grupy skorup białych jak i brązowych, oraz odpowiednio 97% i 95% dla skorup białych i brązowych podczas wykorzystania danych będących transmitancją.

4. * Manualna selekcja kanałów spektralnych i dalsza ich analiza ze względu na specyfikę wykresów transmitancji

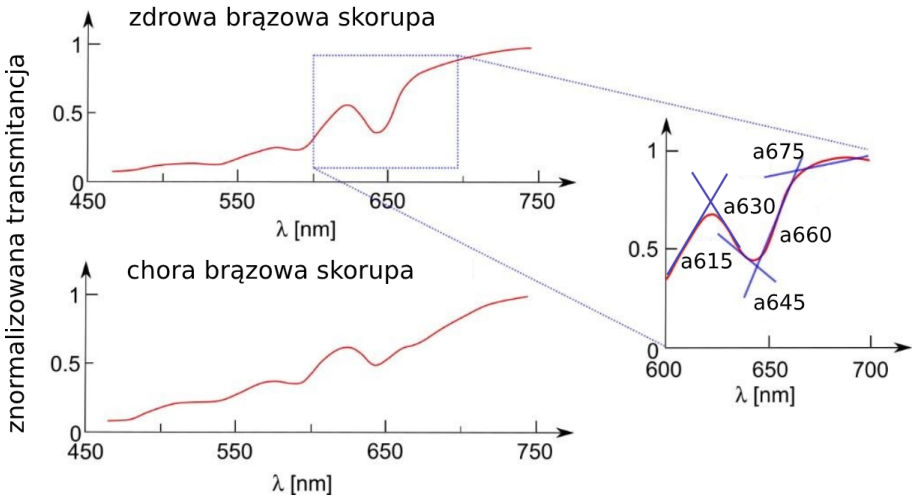


Rysunek 4.6: Zależność średniej transmitancji oraz rozstępu międzykwartylowego IQR zarejestrowanych grup: czerwony - $chorych$ brązowych skorup, zielony – $zdrowych$ brązowych skorup, morski – $chorych$ białych skorup, fioletowy – $zdrowych$ białych skorup, w zależności od długości fali.

W przypadku manualnej selekcji znaczących dla procesu klasyfikacji kanałów spektralnych, bardzo ważna jest odpowiednia wizualizacja zarejestrowanych sygnałów. Zauważalne różnice transmitancji przebadanych grup próbek przedstawia wykres 4.6. Można z niego odczytać relacje średniej transmitancji oraz zakres międzykwartylowy (IQR) wyników danej grupy w zależności od długości fali. Wyszczególniony na żółto obszar oznacza największe zróżnicowanie geometrii wykresów między sobą, wskazując jednocześnie zawężony obszar (600 – 700 nm), do którego zostanie ograniczona dalsza, manualna część analizy danych.

Po przetestowaniu kilku ręcznie wyselekcjonowanych parametrów a będących efektem dopasowania prostych do wizualnie różniących się fragmentów wykresu transmitancji uzyskano acc o wartości 88% dla grupy skorup brązowych i 84% dla grupy skorup białych. Najlepsze wyniki przyporządkowania próbek do klas uzyskano dla parametrów $zakres$ = $skok$ = 15 nm. Poglądowe przedstawienie dopasowanych prostych wraz z oznaczeniami parametrów a dla przypadku z grupy skorup brązowych, pokazane jest na rysunku 4.7. W badanym zakresie mieści się 5 takich prostych, jednak wykorzystanie w pełni metody parametryzacji umożliwia dostrzeżenie, że klasyfikator wybiera jedynie dwie z nich w przypadku grupy skorup brązowych i trzy w przypadku grupy skorup białych. Wykazują to drzewa decyzyjne wygenerowane dla obu grup (rysunki 4.8 i 4.9).

Dzięki łatwej analizie wyników wygenerowanych przez DT można dostrzec, że w przypadku skorup brązowych wystarczy tylko jedno kryterium podziału ($a675 \leqslant 0.016$), które



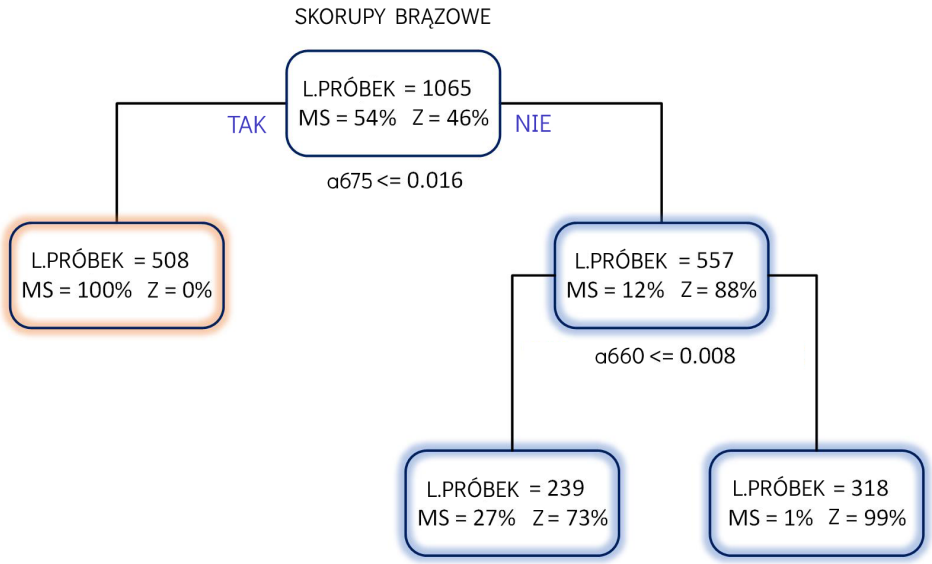
Rysunek 4.7: Poglądowa wizualizacja ręcznego dopasowania wielomianów pierwszego stopnia do wybranego fragmentu wykresu transmitancji na przykładzie skorup $zdrowej$ i $chorej$ z grupy skorup brązowych. Przedział 600–700 nm, $zakres$ = $skok$ = 15 nm. Np. a615 oznacza prostą dopasowaną do zakresu 15 nm od wartości dł. fali = 615 nm.

w węźle głównym dokonuje ze 100% skutecznością wyboru skorup $chorych$ (pozostawiając ciągle część próbek od kur z MS w pozostałym zbiorze). Parametr tak dobrze rozgraniczający nie występuje w grupie skorup białych. Aby uzyskać 88% skuteczności prawidłowej klasyfikacji w zbiorze skorup brązowych należy wykorzystać parametry: a660 i a675, natomiast wynik 84% dla zbioru skorup białych, uzyskuje się wykorzystując: a630, a645 i a675. Oznacza to, że wartość a615 nie wprowadza istotnej informacji do klasyfikatora.

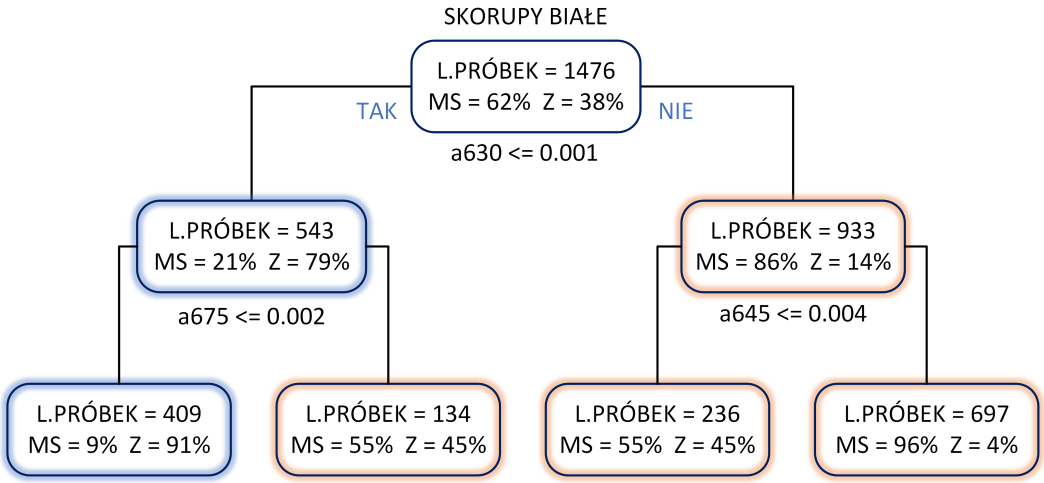
Wyniki klasyfikacji uzyskane różnymi metodami obliczeniowymi zamieszczono w tabeli 4.2. Najwyższą wartość współczynnika dokładności – 97% osiąga się w grupie skorup białych, wykorzystując klasyfikator SVM na danych będących bezpośrednio transmitancją. Tą samą wartość procentową, po 7-krotnej kroswalidacji, można uzyskać w obu grupach skorup stosując metodę PAW zaproponowaną w rozprawie.

Podsumowanie i wnioski

Otrzymana dokładność klasyfikacji próbek na poziomie 95% – 97% jest najwyższą otrzymaną za pomocą przetestowanych algorytmów. Efekt działań zaimplementowanego systemu do klasyfikacji miał za zadanie wskazać istnienie podwyższonego ryzyka występowania kur zakażonych MS w stadzie lub dowieść jego braku, na podstawie analiz spektralnych VIS



Rysunek 4.8: Najlepszy model DT uzyskany z manualnej selekcji kanałów spektralnych dla grupy skorup brązowych. acc = 88% po 7-krotnej kroswalidacji przy ręcznie dobranych parametrach: $zakres$ = $skok$ = 15 nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*.



Rysunek 4.9: Najlepszy model DT uzyskany z manualnej selekcji kanałów spektralnych dla grupy skorup białych. acc = 84% po 7-krotnej kroswalidacji przy ręcznie dobranych parametrach: $zakres$ = $skok$ = 15 nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*.

Tabela 4.2: Podsumowanie wyników klasyfikacji wykonanych za pomocą klasyfikatorów: DT, ANN – typ RBF i SVM na danych w trzech różnych formach. a – wartości współczynników kierunkowych prostych dopasowanych do transmitacji przy wykorzystaniu dobranych parametrów (jeden z efektów działania metody PAW); PCA (3 PC) – 3 pierwsze składowe główne obliczane z transmitancji; PCA (10 PC) – 10 pierwszych składowych głównych obliczanych z transmitancji; manual. s.* – manualna selekcja kanałów spektralnych; acc [%] – dokładność DT po 7-krotnej kroswalidacji (7-k. kw.); $acc\ r.$ [%] – współczynnik dokładności klasyfikatora powstały w wyniku walidacji prostej (w.p.). Wytłuszczono wartości uzyskane w wyniku działania systemu do klasyfikacji zaproponowanego przez autorkę.

dane wejściowe	klasyfikator	białe skorupy		brązowe skorupy	
		acc [%] (7-k. kw.)	$acc\ r.$ [%] (w. p.)	acc [%] (7-k. kw.)	$acc\ r.$ [%] (w. p.)
a	DT /PAW/	**97**	**96**	**97**	**95**
PCA (3 PC)	DT	96	95	88	95
manual. s.*	DT	84	83	88	87
		$acc\ r.$ [%]			
PCA (3 PC)	RBF	95		92	
PCA (10 PC)	RBF	94		85	
PCA (3 PC)	SVM	96		96	
transm.	SVM	97		95	

skorup jaj pochodzących od tych zwierząt. W przypadku pozytywnych testów, konkretne stado miało zostać poddane bardziej szczegółowym badaniom biochemicznym. Wymagania stawiane ww. zadaniu nie wymuszały podwyższania uzyskanych poziomów acc, z tego względu nie zastosowano etapu ponownej redukcji danych widmowych metodą DRW.

Manualna selekcja kanałów spektralnych bazująca na wizualnej analizie wykresu transmitacji i dobraniu parametrów, które umożliwiły wygenerowanie odpowiednich a pozwoliła na uzyskanie dokładności rzędu 88%. Wynik ten jest prawie o 10 punktów procentowych mniejszy niż wynik uzyskany za pomocą metody PAW, jednak i tak wartość prawidłowego przyporządkowania ciągle plasuje się powyżej 80%.

Badania dowiodły, że stosowanie SVM jak i DT na zredukowanych danych za pomocą PCA, do klasyfikacji skorup na *chore* i *zdrowe*, daje zbliżone rezultaty do efektów autorskiej metody parametryzacji. Należy jednak zwrócić uwagę, że wykorzystanie ANN i SVM nie daje możliwości wglądu w strukturę klasyfikacji, natomiast redukcja z wykorzystaniem PCA, powoduje przeniesienie danych do nieinterpretowalnych wymiarów. W takim przypadku dostosowanie układu do konkretnego zadania staje się niewykonalne.

Inaczej jest w przypadku systemu do klasyfikacji zaproponowanego w rozprawie, bazującego na metodzie PAW. Automatycznie dobiera on zakresy spektralne specyficzne dla danego typu próbek, maksymalizując dokładność klasyfikacji przy jednoczesnym nieprzeuczaniu modelu, co przekłada się na możliwość uogólnienia.

W opisywanym przypadku klasyfikacji skorup jaj kurzych na te pochodzące od zainfekowanych kur patogenem MS i na te pochodzące od zdrowych kur analiza grafów drzew decyzyjnych pozwala na konkretną adaptację systemu do zastosowania. Możliwa jest konstrukcja systemu dedykowanego do kontroli stad drobiu pod względem występowania patogenu MS przy zastosowaniu źródeł światła i detektora pracujących jedynie w zakresie 500–720 nm (dokładność klasyfikacji na poziomie 97%). Tak znaczne ograniczenie wymagania co do szerokości rejestrowanego spektrum pozwala na zmniejszenie kosztów produkcji danego systemu, co może spowodować upowszechnienie systemu na dużą skalę.

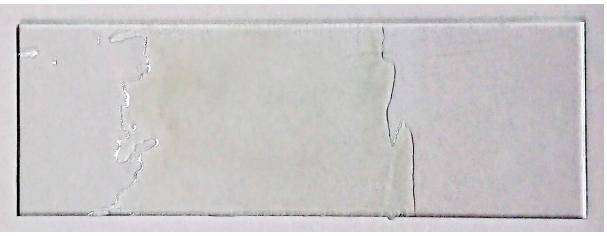
4.2.2. Klasyfikacja miodów ze względu na pochodzenie botaniczne

Wstęp

Miód jest naturalnym produktem pszczelim. Jako produkt spożywczy i handlowy podlega normom zawartym w Rozporządzeniu Ministra Rolnictwa i Rozwoju Wsi z dn. 14 stycznia 2009 roku w sprawie metod analiz związanych z oceną miodu (Dz. U. Nr 17, poz 94) z kolejnymi zmianami oraz PN-88/A-77626 „Miód pszczeli". Badania laboratoryjne miodu, to przede wszystkim badania chemiczne polegające na analizie obecnych w nim pyłków, cukrów (np. analiza chromatograficzna), oznaczeniu aktywności i stężeń enzymów (np. analiza spektrofotometryczna), oznaczeniu przewodności właściwej miodu (metoda konduktometryczna) oraz analizie sensorycznej. Miód może być zafałszowany przez dodanie syropu cukrowego lub pyłku kwiatowego i zanieczyszczony przez obecność pestycydów, środków do zwalczania pasożytów pszczół, antybiotyków, metali ciężkich i drobnoustrojów takich jak bakterie, pleśnie czy drożdże [144].

Typowym sposobem określenia odmian miodów jest analiza pyłkowa, która pozwala na poznanie ich botanicznego pochodzenia. Polega ona na wizualnej ocenie obrazu mikroskopowego. Rozróżnienie cech specyficznych dla pyłków poszczególnych roślin (tj. rozmiar, kształt, złożoność i cechy charakterystyczne) składa się na wiedzę doświadczonego analityka. Na podstawie obrazu pyłkowego miodu i wiedzy o występowaniu roślin charakterystycznych dla danego regionu, kraju, kontynentu czy strefy klimatycznej można rozpoznać pochodzenie miodu, jednak wymaga to dużego doświadczenia badacza [145].

W dalszej części rozdziału przedstawiono system do klasyfikacji próbek miodów ze względu na pochodzenie botaniczne, który po adaptacji do konkretnego zastosowania, konkuruje z metodą pyłkową pod względem szybkości pomiaru oraz zautomatyzowania procesu. W przypadku posiadania małej ilości badanej substancji, co przekłada się na liczbę próbek, zastosowanie opracowanego systemu prowadzi do uzyskania znacząco lepszych wyników klasyfikacji, w porównaniu do użycia typowych metod klasyfikacyjnych na tych samych danych.



Rysunek 4.10: Fotografia przykładowej próbki miodu przygotowanej do pomiaru.

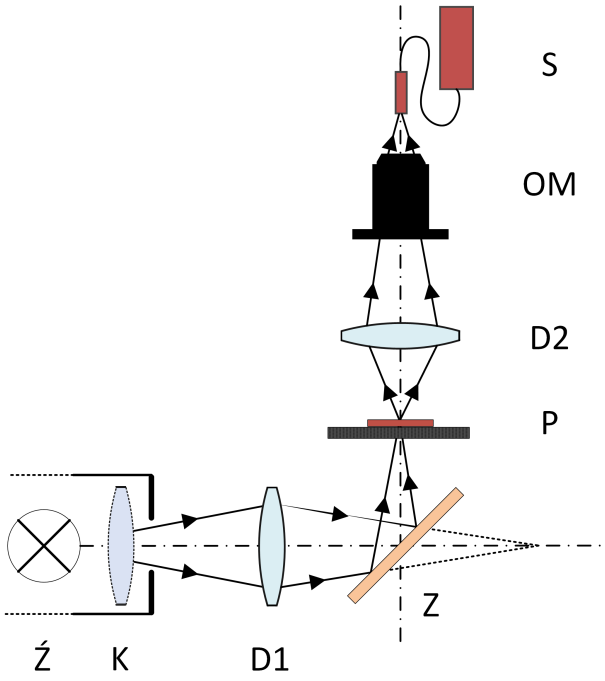
Materiał badawczy

Cztery gatunki miodu – akacjowy (*Robinia pseudoacacia*), gryczany (*Fagopyrum esculentum*), lipowy (*Tilia*) i rzepakowy (*Brassica napus*) zostały poddane klasyfikacji ze względu na pochodzenie botaniczne podczas eksperymentu. Próbki dostarczono z akredytowanego laboratorium Instytutu Ogrodnictwa (Puławy), gdzie dokonano analizy palinologicznej materiału badawczego. Przeprowadzono ją zgodnie z metodologią opisaną w Rozporządzeniu Ministra Rolnictwa i Rozwoju Wsi z dnia 14 stycznia 2009 r. oraz z procedurami zalecanymi przez Międzynarodową Komisję ds. Miodu. Udział pyłku w badanych próbkach przedstawiał się następująco: akacjowy 33%, gryczany 65%, lipowy 54%, rzepakowy 81%.

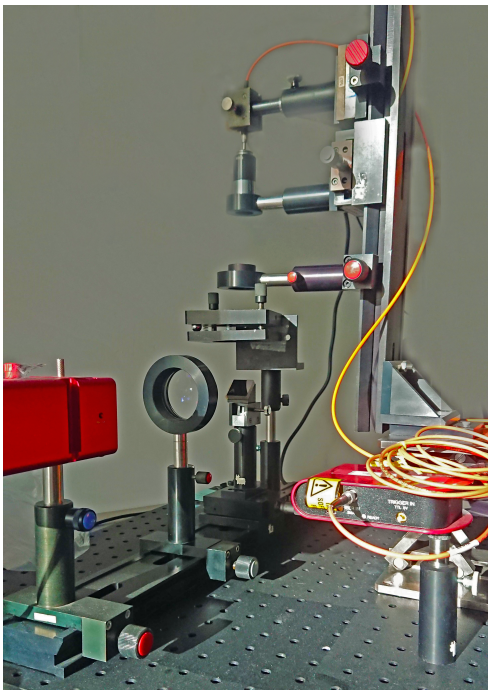
Próbki były przechowywane w specjalnych opakowaniach izolujących od penetracji światła w temperaturze pokojowej przez tydzień. W celu ujednolicenia próbek ze względu na występowanie skrystalizowanego cukru, podgrzano je w specjalistycznym piecu do temperatury $35°\mathrm{C}$ i utrzymano w takowej w przeciągu 5 godzin. Następnie cienka warstwa miodu została nałożona na mikroskopowe szkiełka bazowe. Przykładowa próbka zaprezentowana jest na rysunku 4.10.

Układ optyczny

Próbki z naniesioną warstwą miodu, ze względu na płynną konsystencję materiału badanego nie mogły znajdować się w pozycji pionowej. Z tego powodu układ optyczny z rozdziału 4.2.1 został przerobiony do wersji pionowej tak, żeby próbka mogła być umieszczona w poziomie (schemat, fotografia oraz model 3D układu znajdują się na rysunku 4.11). Dzięki temu materiał nie zmieniał swojej geometrii podczas pomiaru. Zasada działania układu w świetle przechodzącym opisana jest w rozdziale 2.2.2.1. Poszczególne elementy mają analogiczne zadania do tych wykorzystanych w układzie do pomiaru skorup jaj – rozdział 4.2.1.



(a) Schemat optyczny



(b) Fotografia rzeczywistego układu pomiarowego

(c) Rysunek poglądowy

Rysunek 4.11: Schemat (a), fotografia (b) oraz rysunek poglądowy 3D (c) pionowego układu optycznego do pomiarów transmitancji miodów. Ź – stabilizowane inkadescencyjne źródło światła, K – kolektor, D1 i D2 – dublety achromatyczne (o ogniskowych odpowiednio 80 mm i 35 mm), Z – zwierciadło, P – próbka miodu umieszczona na szkiełku bazowym, OM – obiektyw mikroskopowy, S – kompaktowy spektrometr z wejściem światłowodowym.

Pomiary

Kompaktowy spektrometr siatkowy o rozdzielczości widmowej 0.1 nm, pracujący w zakresie VIS (350 nm–750 nm) rejestruje intensywność wiązki światła po przejściu przez warstwę miodu naniesioną na bazę. Wykonano serie pomiarów punktowych z kilku obszarów każdej próbki. Sumarycznie zarejestrowano 60 widm, po 15 każdego rodzaju miodu.

Anliza i wyniki

Podrozdział podzielono na sekcję opisującą analizę i wyniki klasyfikacji z wykorzystaniem metody PAW i DRW oraz sekcję prezentującą rezultaty wykonania obliczeń porównawczych stosując PCA, RBF, SVM. Do przeprowadzenia obliczeń z użyciem niniejszych algorytmów wykorzystano transmitancje widmowe z początkowej sekcji podrozdziału oraz we wskazanych przypadkach zbiory wartości a również obliczone w pierwszej części. Podsumowanie w formie tabelarycznej umieszczono na końcu podrozdziału.

Analiza realizacji wykorzystania systemu z użyciem metody PAW i DRW

Zarejestrowane sygnały intensywnościowe zostały poddane następującej procedurze:

1. obliczenie transmitancji materiału;
   - rejestracja próbek z naniesionym materiałem badanym oraz dodatkowe pomiary czystego szkiełka bazowego, konieczne do obliczenia transmitancji miodu;
2. zastosowanie metody PAW;
   - obliczenie dokładności acc na podstawie DT z kroswalidacją;
3. zastosowanie metody DRW;
   - obliczenie dokładności acc na podstawie DT z kroswalidacją;
   - sprawdzenie – obliczenie współczynnika dokładności $acc\ r.$ za pomocą walidacji prostej.

Wyniki działania systemu z użyciem metody PAW i DRW

Zastosowanie metody PAW pozwoliło na osiągnięcie wyniku acc = 89% przy dobraniu parametrów $zakres$ = $skok$ = 1 nm. Porównując rezultat z innymi pracami poruszającymi zagadnienie pomiarów spektralnych w klasyfikacji miodów (tabela 4.3) zdecydowano się na

Tabela 4.3: Uzyskane wartości dokładności klasyfikacji miodów przez inne grupy naukowe. Badania wykonywane na rożnych zakresach spektralnych. PC-NBC – dwie pierwsze składowe główne z PCA zastosowane w naiwnym klasyfikatorze Bayesa; PC-KMC – dwie pierwsze składowe główne z PCA zastosowane w metodzie k najbliższych sąsiadów; ANN – sztuczne sieci neuronowe; SVM – maszyna wektorów nośnych; LDA – liniowa analiza dyskryminacyjna; SIMCA – proste modelowanie analogii klas; GA-SVM – algorytm genetyczny wykorzystujący SVM; PCA-SVM – składowe główne z PCA zastosowane w metodzie SVM.

acc	metoda	publikacja
88.6 %	PC-NBC	[29]
77.1 %	PC-KMC	
95 %	ANN	[146]
92 %	SVM	
90 %	LDA	
65.625 %	SIMCA	[147]
87.5 %	GA-SVM	
90.62 %	PCA-SVM	

wykonanie ponownego zmniejszenia wymiarowości za pomocą metody doboru i redukcji danych widmowych.

Wyniki testów kilku szerokości okien wycinających powstałych przy zastosowaniu metody DRW zaprezentowane zostały na wykresie 3.8. Najwyższe acc osiągane jest dla okna o szerokości 200 nm umiejscowionego tak, że jego początek odpowiada długości fali 500 nm. Wartość acc obliczona na danych, które pozostały po redukcji, wyniosła 95%.

Wykonanie obliczeń walidacji prostej na danych uzyskanych dzięki DRW poskutkowało osiągnięciem wartości $acc\ r.$ równej 100%. Liczność zbioru uczącego w stosunku do zbioru testowego wynosiła 4:1.

Analiza i wyniki działania innych algorytmów

1. PCA z danych spektralnych, obliczenie acc i $acc\ r.$

   Wykonanie PCA na danych spektralnych i wprowadzenie 3 pierwszych składowych głównych do klasyfikatora DT z 7-krotną kroswalidacją umożliwiło uzyskanie 47% prawidłowej klasyfikacji. Walidacja prosta pozwoliła uzyskać $acc\ r.$ równe jedynie 33%. Otrzymanie tak niskich wartości dokładności poskutkowało sprawdzeniem efektów przeprowadzenia analogicznych obliczeń z wykorzystaniem wcześniej wygenerowanych danych – a.

2. * PCA na wcześniej wygenerowanych danych – a, obliczenie acc i $acc\ r$.

   Przeprowadzenie analogicznej do punktu 1. procedury obliczeniowej na wartościach a dało znacznie lepsze rezultaty: acc – 87%. Jednak wynik walidacji prostej nie potwierdza wysokiej wiarygodności tego typu klasyfikacji ($acc\ r.$ = 58%).

   Zróżnicowanie wyników pomiędzy zastosowaniem algorytmu PCA na danych transmitancji i zbiorze a jest widoczne na wykresach opisujących wariancję za pomocą pierwszej – PC1, drugiej – PC2 i trzeciej składowej głównej – PC3 (rysunek 4.12). W przypadku danych w postaci bezpośredniej transmitancji (podpunkty a i b rysunku 4.12) zarówno zależności PC1 od PC3, jak i PC2 od PC3, nie wykazują potencjału mogącego przełożyć się na pozytywny wynik klasyfikacji (klasy poszczególnych miodów znacznie na siebie nachodzą). Analogiczne zależności składowych głównych obliczonych na bazie a (podpunkty c i d rysunku 4.12) wykazują większe możliwości separacji danych (poszczególne klasy są lepiej rozgraniczone).

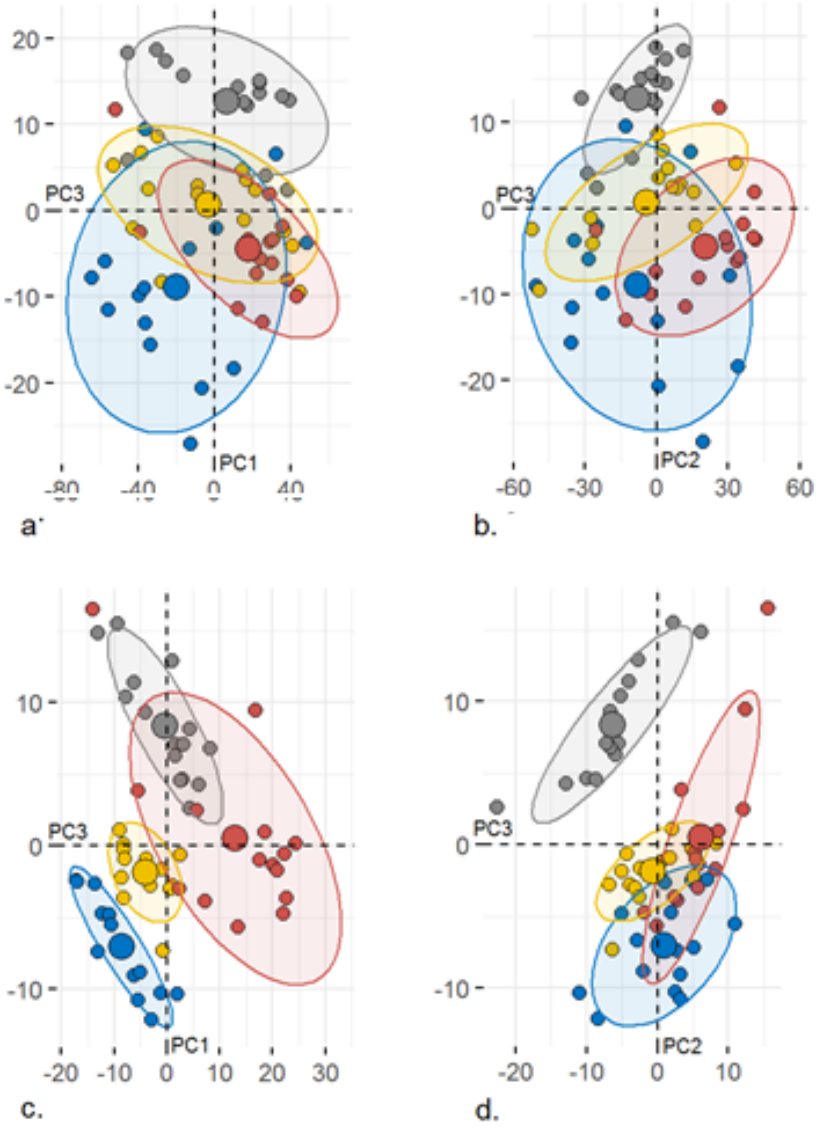
3. RBF na 3 i 10 składowych głównych PCA

   Po wprowadzeniu 3 pierwszych składowych głównych transmitancji do ANN (typ RBF) uzyskano współczynnik dokładności na poziomie 33%. Efekt nie zmienił się po wykorzystaniu 10 pierwszych składowych PC. Dobrane parametry i cechy sieci do eksperymentu dotyczącego skorup jaj kurzych, zostały nie zmienione.

4. SVM na 3 składowych głównych PCA oraz na danych będących transmitancją

   Podczas implementacji SVM na danych o zmniejszonej wymiarowości za pomocą PCA posłużono się modelem maszyny wektorów nośnych analogicznym do tego przedstawionego w poprzednim eksperymencie. Wykonanie SVM na danych transmitancji poskutkowało osiągnięciem $acc\ r.$ = 22%. Zredukowanie danych za pomocą PCA podniosło wynik do 33%.

Wyniki klasyfikacji uzyskane różnymi metodami obliczeniowymi zamieszono w tabeli 4.4. Najwyższą wartość współczynnika dokładności osiągnięto stosując metodę PAW w połączeniu z metodą DRW zaproponowaną w rozprawie. Wykorzystanie klasyfikatorów typu ANN i SVM na analizowanych danych nie przyniosło satysfakcjonujących rezultatów.



Rysunek 4.12: Wykresy zależności składowych głównych, pierwszej (PC1) i drugiej (PC2) od trzeciej (PC3). a i b obliczenia na danych transmitancji, b i d obliczenia na danych w postaci zbioru parametrów a. Znaczniki: niebieskie – miód gryczany, żółte – miód lipowy, szare – miód akacjowy (robinia), czerwone – miód rzepakowy.

Tabela 4.4: Podsumowanie wyników klasyfikacji wykonanych za pomocą klasyfikatorów: DT, ANN – typ RBF i SVM na danych w trzech różnych formach. a – wartości współczynników kierunkowych prostych dopasowanych do transmitancji przy wykorzystaniu dobranych parametrów (jeden z efektów działania metody PAW); PCA – 3 pierwsze składowe główne obliczane z transmitancji (wyniki bazujące na 10 pierwszych składowych głównych nie różniły się); PCAa* – 3 pierwsze składowe główne obliczone na zbiorze a; transm. – transmitancja próbek po odszumieniu za pomocą filtra S-G; acc [%] – dokładność DT po 7-krotnej kroswalidacji (7-k. kw.); $acc\ r.$ [%] – współczynnik dokładności klasyfikatora powstały w wyniku walidacji prostej (w.p.). Wytłuszczono wartości uzyskane w wyniku działania systemu do klasyfikacji zaproponowanego przez autorkę.

dane wejściowe	klasyfikator	acc [%] (7-k. kw.)	$acc\ r.$ [%] (w.p.)
a	DT /PAW, DRW/	**95**	**100**
PCA	DT	47	33
PCAa*	DT	87	58
			$acc\ r.$ [%]
PCA	RBF		33
PCA	SVM		33
transm.	SVM		22

Podsumowanie i wnioski

Porównując wyniki efektu analizy tych samych danych spektralnych za pomocą przetestowanych w rozprawie powszechnie znanych metod klasyfikacyjnych oraz metod PAW i DRW, można zauważyć znaczącą różnicę w poziomach dokładności klasyfikacji. Główną przyczyną niskich wartości $acc\ r.$ dla ANN (z wykorzystaniem RBF) i SVM prawdopodobnie jest zbyt mała liczba próbek. Algorytmy te wymagają do prawidłowego działania liczby próbek co najmniej równej liczbie wymiarów danych. W omawianym przypadku to kryterium nie było możliwe do spełnienia.

Wykonanie PCA na danych transmitancji dało gorsze wyniki, niż na danych w postaci parametrów a. Jest to widoczne zarówno na wykresach zależności składowych głównych PC, jak i w obliczosnych wartościach acc i $acc\ r.$ Fakt ten dowodzi, że metody parametryzacji pozytywnie oddziałują na klasyfikację.

Uzyskany podczas wykorzystania metod parametryzacji i redukcji wynik prawidłowej klasyfikacji na poziomie 95%, jest lepszy w porównaniu do rezultatów prezentowanych przez inne zespoły naukowe (tabela 4.3). Żadna z wymienionych grup nie klasyfikowała miodów wyłącznie bazując na widmie VIS.

Dostosowując system do konkretnego zastosowania, w tym przypadku stosując np. doświetlenie próbki w przedziale 350–500 nm (zakres wskazany jest na podstawie wyniku dotyczącego umiejscowienia okna wycinającego), możliwe jest uzyskanie taniego i praktycznego urządzenia do klasyfikacji 4 typów omawianych w badaniu miodów pod względem pochodzenia botanicznego – dokładność klasyfikacji na poziomie 95%. Polepszenie wyników po wycięciu fragmentu spektrum może świadczyć o tym, że występuje czynnik wpływający w ten sam sposób na wszystkie badane typy miodów w określonym zakresie widma. Dlatego pozostawienie go w zbiorze danych tworzących model klasyfikacyjny, może pogorszyć wynik dokładności klasyfikacji. Przykładem takiego czynnika, mogą być różnego typu zanieczyszczenia miodów wpływające na widmo w zakresie 500–700 nm.

W takim zakresie zaproponowany w rozprawie system staje się konkurencyjny w stosunku do analizy pyłkowej, będąc metodą mniej czasochłonną i niewymagającą pogłębionej wiedzy palinologicznej powiązanej z dużym doświadczeniem analitycznym badacza, co powoduje, że metoda staje się również atrakcyjna ekonomicznie.

4.3. Podsumowanie rozdziału

W rozdziale zaprezentowano dwa przeprowadzone eksperymenty. Badania wykazały potencjał wykorzystania wysokorozdzielczych pomiarów operujących jedynie na zakresie VIS w celach klasyfikacji obiektów warstwowych. Uzyskane wyniki były porównywalne z wartościami publikowanymi przez inne grupy badawcze, a czasem lepsze. Zarówno w przypadku badania skorup jaj kurzych, jak i miodów, proces ujednolicenia grubości próbek jest bardzo utrudniony. Dzięki zastosowaniu metody parametryzacji podczas użycia systemu stosującego PAW niewielkie zróżnicowanie grubości badanych obiektów warstwowych nie stanowiło problemu. W obu przedstawionych przypadkach przeprowadzona analiza doprowadziła do możliwości przystosowania systemu do konkretnego zastosowania. Wszyscy przytoczeni w pracy badacze wykorzystywali szerszy zakres spektralny do analiz niż VIS. Jest to jednoznacznie związane ze stwierdzeniem, że opracowany system do klasyfikacji obiektów warstwowych wykorzystujący techniki spektralne VIS jest znacznie bardziej atrakcyjny ekonomicznie w stosunku do systemów badawczych zaprezentowanych przez wymienione grupy badawcze.

5. Podsumowanie rozprawy

5.1. Wnioski, podsumowanie rozprawy oraz kierunki dalszych prac

Niniejsza rozprawa doktorska kompleksowo opisuje zagadnienie klasyfikacji obiektów warstwowych przy wykorzystaniu technik spektralnych, skupiając się na zakresie promieniowania widzialnego.

W pracy przedstawiono podstawy fizycznych zjawisk zachodzących w materii podczas rejestracji spektralnych, zawarto najważniejsze – zdaniem autorki – wybrane zagadnienia przetwarzania danych intensywnościowych oraz klasyfikacji. Wskazano ograniczenia współczesnych systemów oraz zaproponowano rozwiązanie mogące w konkretnych sytuacjach konkurować z istniejącymi metodami. Rozprawa zawiera szczegółowy opis dwóch eksperymentów wskazujących obszary potencjalnych zastosowań urządzenia.

Praca wpisuje się dziedzinę spektroskopii, przedstawiając możliwości, jakie niesie ze sobą wykorzystanie jedynie zakresu promieniowania VIS. Przedstawia punktową technikę pomiarową umożliwiającą dokonanie klasyfikacji mało licznego zbioru próbek, nie poddanych uprzednio specjalnemu przygotowaniu. Algorytmy typu ANN, jak również SVM, mogą sobie nie radzić z zagadnieniem małej liczebności próby, co przedstawiono w pracy. Można stwierdzić, iż prezentowany system łączy w sobie punktowość pomiarów spektroskopowych z pewnego rodzaju uogólnieniem interpretacji sygnału występującym w obrazowaniu spektralnym. Opracowany system charakteryzuje się dużą elastycznością i możliwością dostosowania do konkretnego przypadku.

Głównym czynnikiem wyróżniającym zaproponowany w rozprawie system w stosunku do wykorzystywanych w dziedzinie jest zastosowanie zakresu spektralnego obejmującego światło widzialne umożliwiające klasyfikację wybranych obiektów z dokładnością powyżej 80%. Wyniki możliwe były do uzyskania dzięki zastosowaniu dwóch opracowanych metod analizy danych spektralnych: metody parametryzacji z użyciem aproksymacji wielomianowej (PAW) oraz doboru i redukcji widma (DRW). Osiągnięte efekty były porównywalne bądź lepsze

w stosunku do tych uzyskanych po implementacji przetestowanych w pracy, powszechnie znanych algorytmów. Za pomocą metody PAW generowane są nowe cechy sygnałów wielowymiarowych niosących ze sobą informację o konkretnych zakresach widmowych. Dzięki tak zaprojektowanej parametryzacji możliwa jest dalsza interpretacja uzyskanych wyników. Działanie metody DRW ma za zadanie redukcję widma o mało znaczący dla konkretnego problemu klasyfikacyjnego fragment, po usunięciu którego dokładność klasyfikacji wzrasta. Zawężenie obszaru pomiarowego w kontrolowany sposób pozwala na świadomy dobór elementów optycznych pracujących w ograniczonych zakresach. Adaptacja systemu wiąże się z obniżeniem kosztów jego produkcji, a co za tym idzie z możliwością upowszechnienia. Dopuszczająca modyfikacje mechanika systemu optycznego jest nieskomplikowana, co sprawia, że staje się on niezawodny. Popularyzacja tego typu rozwiązań wydaje się być realna tym bardziej, że urządzenia nie zawierają niebezpiecznych elementów, praca odbywa się w zakresie spektralnym bezpiecznym dla człowieka, układy są nieskomplikowane i mobilne, a oprogramowanie umożliwia zautomatyzowanie procesu.

W pracy zaprezentowano system będący w konfiguracji realizującej pomiary punktowe, jednak możliwe jest przekształcenie go w układ skanujący. Przykładem wykorzystania systemu w ustawieniu dostosowanym do akwizycji pomiarów powierzchniowych są badania właściwości optycznych past grafenowych [10, 148], którymi również zajmowała się autorka rozprawy.

Postawiony cel pracy w postaci opracowania systemu wykorzystującego pomiary widma promieniowania z zakresu widzianego wraz z implementacją metod parametryzacji i redukcji sygnału do klasyfikacji obiektów warstwowych został osiągnięty.

Dalsze prace rozwijające przedstawione zagadnienie będą skupiać się na optymalizacji stworzonych algorytmów, w szczególności do zastosowań w pomiarach polowych. Znaczącym etapem rozwoju systemu stanie się również testowanie możliwości klasyfikacji różnych obiektów i określenie typu materiałów oraz zakresu badań możliwych do przeprowadzenia za pomocą opisanego w rozprawie systemu.

5.2. Elementy nowości w pracy

W rozprawie zostały przedstawione następujące zagadnienia, nieopisane wcześniej w literaturze:

- metoda parametryzacji sygnału widmowego z użyciem aproksymacji wielomianowej (PAW) – podrozdział 3.2.1;
- metoda doboru i redukcji widma (DRW) – podrozdział 3.2.2;
- opracowanie systemu wykorzystującego widmo zakresu VIS do klasyfikacji obiektów warstwowych – podrozdziały 4.1 i 4.2.

W zaprezentowanych przypadkach zastosowanie systemu rozwiązuje problemy związane z:

- niewystarczającą liczbą próbek do skutecznego nauczenia sieci neuronowych (w szczególności w pomiarach wysokowymiarowych) – wykorzystanie w algorytmach metody PAW, lub metod PAW i DRW;
- trudną dostępnością próbek – mobilność systemu;
- niepowtarzalnością wymiaru w osi z badanych próbek – parametryzacja wykresu transmitancji uniezależniająca wynik od względnej zarejestrowanej intensywności.

Spis rysunków

2.1 Schemat procesu klasyfikacji obiektów warstwowych wykorzystujący techniki spektralne VIS. „Char." – charakterystyki. 27

2.2 Schematy układów do pomiaru (a) reflektancji, (b) transmitancji próbki, bez wykorzystania bazy. I_{zr} – intensywność światła docierającego do próbki, I_p – intensywność zarejestrowana po odbiciu lub przepuszczeniu przez materiał badany. 29

2.3 Schemat układu do pomiaru (a) transmitancji próbki wraz z bazą, (b) intensywności wiązki po przejściu przez bazę – bez nałożonego materiału testowanego. I_{zr} – intensywność światła docierającego do bazy, I_{sum} – sumaryczna intensywność zarejestrowana po przejściu przez bazę i materiał badany, I_b – intensywność zarejestrowana po przejściu przez samą bazę. 30

2.4 Graficzna wizualizacja idei binarnego drzewa decyzyjnego. 41

2.5 Graficzna wizualizacja modelu neuronu. 44

2.6 Graficzna wizualizacja modelu sieci RBF. 45

2.7 Graficzna wizualizacja zasady działania SVM dla danych liniowo separowalnych. 47

3.1 Graficzna wizualizacja parametrów: $zakres$ i $skok$ wraz z aproksymacją wielomianów stopnia 1. metodą najmniejszych kwadratów. Tangensy kątów nachylenia prostych zaznaczonych na czerwono są ich współczynnikami kierunkowymi oznaczonymi $a\lambda_{11}$ i $a\lambda_{21}$. 53

3.2 Schemat metody PAW – parametryzacji z użyciem aproksymacji wielomianowej oraz metody DRW – doboru i redukcji widma. Zielona ramka – wynik konkretnego etapu, szary prostokąt – okno wycinające o zadanej szerokości. 55

3.3 Wartości dokładności acc w zależności od kombinacji parametrów $zakres$ i $skok$, z uwzględnieniem uszeregowania $skoku$ w sposób rosnący. Szara linia – acc [%]; zielona linia – aproksymacja za pomocą średniej ruchomej acc [%]; niebieskie słupki – $zakres$ [nm]; pomarańczowe słupki – $skok$ [nm]. Wykres należy czytać w następujący sposób: dla $zakresu$ x nm i $skoku$ y nm wartość acc wynosi z%. Oś x przedstawia kolejne kombinacje parametrów $zakres$ i $skok$. 58

3.4 Wartości dokładności acc w zależności od kombinacji parametrów $zakres$ i $skok$, z uwzględnieniem uszeregowania $zakresu$ w sposób rosnący. Szara linia – acc [%]; zielona linia – aproksymacja liniowa acc [%]; niebieskie słupki – $zakres$ [nm]; pomarańczowe słupki – $skok$ [nm]. Wykres należy czytać w następujący sposób: dla $zakresu$ x nm i $skoku$ y nm wartość acc wynosi z%. Oś x przedstawia kolejne kombinacje parametrów $zakres$ i $skok$. 59

3.5 Kombinacje parametrów $zakres$ i $skok$ w wyniku zastosowania których uzyskuje się dokładności acc powyżej 80%. Szara linia – acc [%]; niebieskie słupki – $zakres$ [nm]; pomarańczowe słupki – $skok$ [nm]. Wykres należy czytać w następujący sposób: dla $zakresu$ x nm i $skoku$ y nm wartość acc wynosi z%. Oś x przedstawia kolejne kombinacje parametrów $zakres$ i $skok$. 60

3.6 Zależność uśrednionego z 50-ciu powtórzeń acc obliczonego na podstawie sparametryzowanych danych z pełnego zakresu spektralnego, w stosunku do kombinacji parametrów o zrównanych wartościach względem siebie. Czerwona kropkowana linia – aproksymacja wielomianowa wartości acc. 61

3.7 Schematyczne przedstawienie idei działania okna wycinającego na danych będących transmitancją sygnału. Tu: okno o szerokości 100 nm, zakres 540–640 nm jest usuwany. . . 62

3.8 Wartości dokładności acc obliczone po usunięciu odpowiednio umiejscowionych okien o szerokościach: 40 nm – niebieskie kropki, 60 nm – pomarańczowe kropki, 100 nm – szare kropki i 200 nm – żółte kropki, w zależności od położenia okna wycinającego na wykresie transmitancji (por. 3.7) (wartość długości fali odpowiada początkowej pozycji okna). Czarna linia – uśrednione ze stu powtórzeń acc obliczone na pełnym zakresie widma. 63

4.1 Schemat idei działania opracowanego systemu do klasyfikacji obiektów warstwowych wykorzystującego techniki spektralne VIS. 73

4.2 Przykłady analizowanych próbek skorup jaj kurzych: (a) brązowe $zdrowe$ skorupy; (b) białe $zdrowe$ skorupy; (c) brązowe $chore$ skorupy; (d) białe $chore$ skorupy; (e) $chora$ skorupa bez widocznych deformacji [140]. 77

4.3 Schemat układu optycznego do pomiarów transmitancji skorup jaj, bez zachowanej skali. Ź – stabilizowane inkadescencyjne źródło światła, K – kolektor, Φ_{pp} – średnica otworu przysłony polowej, D – dublet achromatyczny, jego oprawa pełni rolę źrenicy wejściowej będąc jednocześnie przysłoną aperturową (Φ_{pa}), f'$_d$ – ogniskowa dubletu, P – próbka (skorupa jaja kurzego) z zaznaczonym na pomarańczowo obszarem pomiarowym o średnicy Φ_{pl} plamki światła, OM – obiektyw mikroskopowy, f'$_{om}$ – ogniskowa soczewki symbolizującej obiektyw mikroskopowy, S – kompaktowy spektrometr z wejściem światłowodowym, Φ_w – średnica wejściowa światłowodu, s$_d$ – odległość dubletu od przysłony pola, s'$_d$ – odległość próbki od dubletu, s$_{om}$ – odległość próbki od obiektywu, s'$_{om}$ – odległość czoła światłowodu od obiektywu. 78

4.4 Wynik działania metody PAW w postaci drzewa decyzyjnego dla grupy skorup brązowych. acc = 97% po 7-krotnej kroswalidacji przy automatycznie dobranych parametrach: $zakres$ = $skok$ = 5 nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*. 82

4.5 Wynik działania metody PAW w postaci drzewa decyzyjnego dla grupy skorup białych. acc = 97% po 7-krotnej kroswalidacji przy automatycznie dobranych parametrach: $zakres$ = $skok$ = 1 nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*. 82

4.6 Zależność średniej transmitancji oraz rozstępu międzykwartylowego IQR zarejestrowanych grup: czerwony - $chorych$ brązowych skorup, zielony – $zdrowych$ brązowych skorup, morski – $chorych$ białych skorup, fioletowy – $zdrowych$ białych skorup, w zależności od długości fali. 84

4.7 Poglądowa wizualizacja ręcznego dopasowania wielomianów pierwszego stopnia do wybranego fragmentu wykresu transmitancji na przykładzie skorup $zdrowej$ i $chorej$ z grupy skorup brązowych. Przedział 600–700 nm, $zakres$ = $skok$ = 15 nm. Np. a615 oznacza prostą dopasowaną do zakresu 15 nm od wartości dł. fali = 615 nm. 85

4.8 Najlepszy model DT uzyskany z manualnej selekcji kanałów spektralnych dla grupy skorup brązowych. acc = 88% po 7-krotnej kroswalidacji przy ręcznie dobranych parametrach: $zakres$ = $skok$ = 15 nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*. 86

4.9 Najlepszy model DT uzyskany z manualnej selekcji kanałów spektralnych dla grupy skorup białych. acc = 84% po 7-krotnej kroswalidacji przy ręcznie dobranych parametrach: $zakres$ = $skok$ = 15 nm. MS – procent próbek określonych jako *chore*, Z – procent próbek określonych jako *zdrowe*. 86

4.10 Fotografia przykładowej próbki miodu przygotowanej do pomiaru. 90
4.11 Schemat (a), fotografia (b) oraz rysunek poglądowy 3D (c) pionowego układu optycznego do pomiarów transmitancji miodów. Ź – stabilizowane inkadescencyjne źródło światła, K – kolektor, D1 i D2 – dublety achromatyczne (o ogniskowych odpowiednio 80 mm i 35 mm), Z – zwierciadło, P – próbka miodu umieszczona na szkiełku bazowym, OM – obiektyw mikroskopowy, S – kompaktowy spektrometr z wejściem światłowodowym. 91
4.12 Wykresy zależności składowych głównych, pierwszej (PC1) i drugiej (PC2) od trzeciej (PC3). a i b obliczenia na danych transmitancji, b i d obliczenia na danych w postaci zbioru parametrów a. Znaczniki: niebieskie – miód gryczany, żółte – miód lipowy, szare – miód akacjowy (robinia), czerwone – miód rzepakowy. 95

Spis tabel

2.1 Podsumowanie i porównanie cech podstawowych metod redukcji wymiarowości generujących macierze nowych cech. LDA – liniowa analiza dyskryminacyjna, PCA – analiza składowych głównych, PLS – metoda cząstkowych najmniejszych kwadratów. . . . 35

2.2 Podsumowanie i porównanie cech podstawowych metod automatycznej selekcji widmowej (SFS/SBS – selekcja postępująca/wsteczna i SFFS – ruchoma selekcja postępująca) oraz nieautomatycznej (selekcja ręczna i autorska metoda redukcji). Porównanie cech algorytmów zaprezentowano na przykładzie połączenia ich z klasyfikatorem DT – drzewo decyzyjne. 38

3.1 Wyniki acc uzyskane w symulacji trzech zbiorów danych o odchyłkach od wartości średniej: +/- 2%, +/- 5% +/- 10% przy wykorzystaniu klasyfikacji metodą PAW. Wytłuszczono oryginalny pomiar: +/- 1%. 65

3.2 Wpływ rozdzielczości pomiarowej w dziedzinie spektralnej zamodelowanych zbiorów danych na dokładności klasyfikacji (acc) uzyskane metodą PAW, dla różnych niepewności pomiarowych. Wytłuszczono oryginalny pomiar. 66

3.3 Porównanie wybranych metod selekcji cech z uwzględnieniem redukcji wymiarowości poprzez generację nowych cech i selekcję widmową w połączeniu z wybranymi klasyfikatorami. LDA – liniowa analiza dyskryminacyjna, PCA – analiza składowych głównych, PLS – metoda cząstkowych najmniejszych kwadratów, SFFS – algorytm ruchomej selekcji postępującej, DT – drzewo decyzyjne, ANN – sztuczne sieci neuronowe, SVM – maszyna wektorów nośnych. 70

4.1 Podsumowanie liczby próbek skorup $chorych$ i $zdrowych$ będących składowymi zbioru uczącego i testowego w grupie skorup białych i brązowych. 81

4.2 Podsumowanie wyników klasyfikacji wykonanych za pomocą klasyfikatorów: DT, ANN – typ RBF i SVM na danych w trzech różnych formach. a – wartości współczynników kierunkowych prostych dopasowanych do transmitacji przy wykorzystaniu dobranych parametrów (jeden z efektów działania metody PAW); PCA (3 PC) – 3 pierwsze składowe główne obliczane z transmitancji; PCA (10 PC) – 10 pierwszych składowych głównych obliczanych z transmitancji; manual. s.* – manualna selekcja kanałów spektralnych; acc [%] – dokładność DT po 7-krotnej kroswalidacji (7-k. kw.); $acc\ r.$ [%] – współczynnik dokładności klasyfikatora powstały w wyniku walidacji prostej (w.p.). Wytłuszczono wartości uzyskane w wyniku działania systemu do klasyfikacji zaproponowanego przez autorkę. 87

4.3 Uzyskane wartości dokładności klasyfikacji miodów przez inne grupy naukowe. Badania wykonywane na rożnych zakresach spektralnych. PC-NBC – dwie pierwsze składowe główne z PCA zastosowane w naiwnym klasyfikatorze Bayesa; PC-KMC – dwie pierwsze składowe główne z PCA zastosowane w metodzie k najbliższych sąsiadów; ANN – sztuczne sieci neuronowe; SVM – maszyna wektorów nośnych; LDA – liniowa analiza dyskryminacyjna; SIMCA – proste modelowanie analogii klas; GA-SVM – algorytm genetyczny wykorzystujący SVM; PCA-SVM – składowe główne z PCA zastosowane w metodzie SVM. 93

4.4 Podsumowanie wyników klasyfikacji wykonanych za pomocą klasyfikatorów: DT, ANN – typ RBF i SVM na danych w trzech różnych formach. a – wartości współczynników kierunkowych prostych dopasowanych do transmitancji przy wykorzystaniu dobranych parametrów (jeden z efektów działania metody PAW); PCA – 3 pierwsze składowe główne obliczane z transmitancji (wyniki bazujące na 10 pierwszych składowych głównych nie różniły się); PCAa* – 3 pierwsze składowe główne obliczone na zbiorze a; transm. – transmitancja próbek po odszumieniu za pomocą filtra S-G; acc [%] – dokładność DT po 7-krotnej kroswalidacji (7-k. kw.); $acc\ r.$ [%] – współczynnik dokładności klasyfikatora powstały w wyniku walidacji prostej (w.p.). Wytłuszczono wartości uzyskane w wyniku działania systemu do klasyfikacji zaproponowanego przez autorkę. 96

Bibliografia

[1] GUM. *Słowniczek wybranych terminów i definicji stosowanych w metrologii i probiernictwie PL/EN/PL*. Główny Urząd Miar, 2019.

[2] Max Planck. On the law of distribution of energy in the normal spectrum. *Annalen der physik*, 4(553):1, 1901.

[3] DE McCumber. Einstein relations connecting broadband emission and absorption spectra. *Physical Review*, 136(4A):A954, 1964.

[4] Niels Bohr. The spectra of helium and hydrogen. *Nature*, 92(2295):231, 1913.

[5] Alicja Skrzypek and Joanna Matysiak. Techniki stosowane do identyfikacji związków organicznych. *LAB Laboratoria, Aparatura, Badania*, 16:6–12, 2011.

[6] Hanna Jóźwiak. Wykorzystanie spektroskopii w podczerwieni do identyfikacji wyrobów budowlanych. *Prace Instytutu Techniki Budowlanej*, 35:47–55, 2006.

[7] Cécile Gomez, Raphael A Viscarra Rossel, and Alex B McBratney. Soil organic carbon prediction by hyperspectral remote sensing and field vis-nir spectroscopy: An australian case study. *Geoderma*, 146(3-4):403–411, 2008.

[8] Bogdan Zagajewski. Ocena przydatności sieci neuronowych i danych hiperspektralnych do klasyfikacji roślinności tatr wysokich. *Teledetekcja środowiska*, 43, 2010.

[9] Adam Świtoński, Tomasz Błachowicz, Aleksander Sieroń, and Konrad Wojciechowski. A computer-based imaging system for multispectral inspection of skin cancer. *Przegląd Elektrotechniczny*, 88(12b):107–110, 2012.

[10] Zofia Lorenc, Leszek Salbut, Anna Pakula, Marcin Sloma, Grzegorz Wroblewski, and Malgorzata Jakubowska. Optical measurements of selected properties of nanocomposite layers with graphene and carbon nanotubes fillers. 9132:91320E, may 2014.

[11] Eugene Hecht. Hecht optics. *Addison Wesley*, 997:213–214, 1998.

[12] JOSE Torrent and Vidal Barrón. Diffuse reflectance spectroscopy. *Methods of Soil Analysis Part 5—Mineralogical Methods*, 5:367–385, 2008.

[13] George G Guilbault. *Practical fluorescence*, volume 3. CRC Press, 1990.

[14] Inc Encyclopaedia Britannica et al. *Encyclopaedia britannica*. Encyclopaedia Britannica, Incorporated, 1957.

[15] Ernö Pretsch, Philippe Bühlmann, Christian Affolter, Ernho Pretsch, P Bhuhlmann, and C Affolter. *Structure determination of organic compounds*. Springer, 2000.

[16] Yoshihiro Okui, Seiiku Ito, and Masami Sugiyama. Multi-channel spectral light measuring device, March 20 1990. US Patent 4,909,633.

[17] Grzegorz Mączkowski. *Method for measurement of spectral reflectance of the surface of three-dimensional objects*. PhD thesis, The Institute of Micromechanics and Photonics, 2016.

[18] Raju Shrestha, Alamin Mansouri, and Jon Yngve Hardeberg. Multispectral imaging using a stereo camera: Concept, design and assessment. *EURASIP Journal on Advances in Signal Processing*, 2011(1):57, 2011.

[19] Francisco H Imai. Multi-spectral image acquisition and spectral reconstruction using a trichromatic digital camera system associated with absorption filters. *Munsell Color Science Laboratory Technical Report*, 1998.

[20] Til Aach, Johannes Brauers, and Stephan Helling. Multispectral image acquisition with flash light sources. *Journal of Imaging Science and Technology*, 53(3):31103–1, 2009.

[21] Jon Yngve Hardeberg, Francis JM Schmitt, and Hans Brettel. Multispectral color image capture using a liquid crystal tunable filter. *Optical engineering*, 41(10):2532–2549, 2002.

[22] Pierre Jacquinot. The luminosity of spectrometers with prisms, gratings, or fabry-perot etalons. *JOSA*, 44(10):761–765, 1954.

[23] Chein-I Chang. *Hyperspectral imaging: techniques for spectral detection and classification*, volume 1. Springer Science & Business Media, 2003.

[24] Krzysztof Tutak and Mateusz Pieszko. Wykorzystanie obrazowania spektralnego do identyfikacji materiałów na potrzeby sortowania odpadów. *Archives of Waste Management and Environmental Protection*, 17(4):67–78, 2015.

[25] Michel Delhaye. Rapid scanning raman spectroscopy. *Applied optics*, 7(11):2195–2199, 1968.

[26] William L Barnes, Thomas S Pagano, and Vincent V Salomonson. Prelaunch characteristics of the moderate resolution imaging spectroradiometer (modis) on eos-am1. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4):1088–1100, 1998.

[27] Kuanglin Chao, PM Mehl, and YR Chen. Use of hyper–and multi–spectral imaging for detection of chicken skin tumors. *Applied Engineering in Agriculture*, 18(1):113, 2002.

[28] Xiaoli Li and Yong He. Discriminating varieties of tea plant based on Vis/NIR spectral characteristics and using artificial neural networks. *Biosystems Engineering*, 99(3):313–321, 2008.

[29] Diana Tsankova and Svetla Lekova. Botanical origin-based honey discrimination using vis-nir spectroscopy and statistical cluster analysis. *Journal of Chemical Technology and Metallurgy*, 50(5):638–642, 2015.

[30] Deepalekshmi Ponnamma, Didier Rouxel, and Sabu Thomas. Spectroscopy—introducing the advantages and application areas in polymer nanocomposites. In *Spectroscopy of Polymer Nanocomposites*, pages 1–14. Elsevier, 2016.

[31] Gregory A Carter. Responses of leaf spectral reflectance to plant stress. *American Journal of Botany*, 80(3):239–243, 1993.

[32] Laeiq Ahmad, M Tahir Shah, and Shuhab D Khan. Reflectance spectroscopy and remote sensing data for finding sulfide-bearing alteration zones and mapping geology in gilgit-baltistan, pakistan. *Earth Science Informatics*, 9(1):113–121, 2016.

[33] Roy S Berns, E René de la Rie, et al. The relative importance of surface roughness and refractive index in the effects of varnishes on the appearance of paintings. In *ICOM-CC Triennial Meeting Preprints, Rio de Jeneiro*, pages 211–216, 2002.

[34] Norimichi Tsumura, Yoichi Miyake, and Vladimir Bochko. Spectral color imaging system for estimating spectral reflectance of paint. *Journal of Imaging Science and Technology*, 51(1):70–78, 2007.

[35] Alejandro Ribés Cortés. *Multispectral analysis and spectral reflectance reconstruction of art paintings*. PhD thesis, 2003.

[36] Hiroshi Masuhara, FC De Schryver, N Kitamura, and N Tamai. *Microchemistry: Spectroscopy and chemistry in small domains*. Newnes, 2012.

[37] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[38] Lidija Svecnjak, Dragan Bubalo, Goran Baranović, and Hrvoje Novosel. Optimization of ftir-atr spectroscopy for botanical authentication of unifloral honey types and melissopalynological data prediction. *European Food Research and Technology*,

240(6):1101–1115, 2015.

[39] Lea Lenhardt, Ivana Zeković, Tatjana Dramićanin, Živoslav Tešić, Dušanka Milojković-Opsenica, and Miroslav D Dramićanin. Authentication of the botanical origin of unifloral honey by infrared spectroscopy coupled with support vector machine algorithm. *Physica Scripta*, 2014(T162):014042, 2014.

[40] C Herrero Latorre, RM Peña Crecente, S García Martín, and J Barciela García. A fast chemometric procedure based on nir data for authentication of honey with protected geographical indication. *Food chemistry*, 141(4):3559–3565, 2013.

[41] Xiu Ying Liang, Xiao Yu Li, and Wen Jun Wu. Classification of floral origins of honey by nir and chemometrics. In *Advanced Materials Research*, volume 605, pages 905–909. Trans Tech Publ, 2013.

[42] Kezhi Z Mao. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):629–634, 2004.

[43] Franz Pernkopf and Paul O'Leary. Feature selection for classification using genetic algorithms with a novel encoding. In *International Conference on Computer Analysis of Images and Patterns*, pages 161–168. Springer, 2001.

[44] Lea Lenhardt, Rasmus Bro, Ivana Zeković, Tatjana Dramićanin, and Miroslav D Dramićanin. Fluorescence spectroscopy coupled with parafac and pls da for characterization and classification of honey. *Food Chemistry*, 175:284–291, 2015.

[45] Juan Antonio Fernández Pierna, Ouissam Abbas, Pierre Dardenne, and Vincent Baeten. Discrimination of corsican honey by ft-raman spectroscopy and chemometrics. *Base*, 2011.

[46] Xiangrong Zhu, Shuifang Li, Yang Shan, Zhuoyong Zhang, Gaoyang Li, Donglin Su, and Feng Liu. Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics. *Journal of Food Engineering*, 101(1):92–97, 2010.

[47] Mucahid Mustafa Saritas and Ali Yasar. Performance analysis of ann and naive bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, 7(2):88–91, 2019.

[48] Giles M Foody and Ajay Mathur. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for

classification by a svm. *Remote Sensing of Environment*, 103(2):179–189, 2006.

[49] Seher Gok, Mete Severcan, Erik Goormaghtigh, Irfan Kandemir, and Feride Severcan. Differentiation of anatolian honey samples from different botanical origins by atr-ftir spectroscopy using multivariate analysis. *Food chemistry*, 170:234–240, 2015.

[50] Anthony MC Davies, Branka Radovic, Tom Fearn, and Elke Anklam. A preliminary study on the characterisation of honey by near infrared spectroscopy. *Journal of near infrared spectroscopy*, 10(2):121–135, 2002.

[51] Bart J. Kemps, Flip R. Bamelis, Kristof Mertens, Eddy M. Decuypere, Josse G. de Baerdemaeker, and Bart de Ketelaere. Assessment of embryonic growth in chicken eggs by means of visible transmission spectroscopy. *Biotechnology Progress*, 26(2):512–516, 2010.

[52] David M. Gates, Harry J. Keegan, John C. Schleter, and Victor R. Weidner. Spectral Properties of Plants. *Applied Optics*, 4(1):11, 1965.

[53] T. M. Shafey, M. M. Ghannam, H. A. Al-Batshan, and M. S. Al-Ayed. Effect of pigment intensity and region of eggshell on the spectral transmission of light that passes the eggshell of chickens. *International Journal of Poultry Science*, 2004.

[54] Francis Joseph Baker and Reginald Edward Silverton. *Introduction to medical laboratory technology*. Butterworth-Heinemann, 2014.

[55] Kapil Khanal, Santosh Bhusal, Manoj Karkee, and Qin Zhang. Distinguishing one year and two year old canes of red raspberry plant using spectral reflectance. *IFAC-PapersOnLine*, 51(17):39–44, 2018.

[56] Huimin Sun, Yongfang Dong, Pingli Zhang, Yaoyong Meng, Wei Wen, Nan Li, and Zhiyou Guo. Accurate Age Estimation of Bloodstains Based on Visible Reflectance Spectroscopy and Chemometrics Methods. *IEEE Photonics Journal*, 9(1):1–14, 2017.

[57] John C Lindon, George E Tranter, and David Koppenaal. *Encyclopedia of spectroscopy and spectrometry*. Academic Press, 2016.

[58] Saman A. Mehdizadeh, Saeid Minaei, Nigel H. Hancock, and Mohamad Amir Karimi Torshizi. An intelligent system for egg quality classification based on visible-infrared transmittance spectroscopy. *Information Processing in Agriculture*, 1(2):105–114, 2014.

[59] Yun Li and Haiqing Yang. Honey discrimination using visible and near-infrared spectroscopy. *ISRN Spectroscopy*, 2012, 2012.

[60] DR Heath. Optics and vision. In *Telecommunications Engineer's Reference Book*, pages 7–1. Elsevier, 1993.

[61] Tommy Bergmann, Florian Heinke, and Dirk Labudde. Towards substrate-independent age estimation of blood stains based on dimensionality reduction and k-nearest neighbor classification of absorbance spectroscopic data. *Forensic Science International*, 278:1–8, 2017.

[62] Gregory E Stillman et al. Optoelectronics. In *Reference Data for Engineers*, pages 21–1. Elsevier, 2002.

[63] Harald Martens, SA Jensen, and P Geladi. Multivariate linearity transformation for near-infrared reflectance spectrometry. In *Proceedings of the Nordic symposium on applied statistics*, pages 205–234. Stokkand Forlag Publishers Stavanger, Norway, 1983.

[64] P Geladi, D MacDougall, and H Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied spectroscopy*, 39(3):491–500, 1985.

[65] RJ Barnes, Mewa Singh Dhanoa, and Susan J Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, 43(5):772–777, 1989.

[66] Åsmund Rinnan, Frans van den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*, 28(10):1201–1222, 2009.

[67] K Norris and P Williams. Optimization of mathematical treatments of raw near-infrared signal in the. *Cereal Chem*, 61(2):158–165, 1984.

[68] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[69] A J Owen. Uses of Derivative Spectroscopy. *Spectroscopy*, page 8, 1995.

[70] Kun-peng Zhou, Xu-fang Bai, and Wei-hong Bi. Determination of ethanol content in ethanol-gasoline based on derivative absorption spectrometry and information fusion. *Optoelectronics Letters*, 14(6):442–446, 2018.

[71] Tadao Hakuta, Hideyuki Shinzawa, and Yukihiro Ozaki. Practical method for the detection of tetracyclines in honey by hplc and derivative uv-vis spectra. *Analytical Sciences*, 25(9):1149–1153, 2009.

[72] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University

Press, 2007.

[73] RO Duda, PE Hart, and DG Stork. Pattern classification. 2nd edn wiley. *New York*, 153, 2000.

[74] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[75] Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

[76] Mirosław Krzyśko, Waldemar Wołyński, Tomasz Górecki, and Michał Skorzybut. Systemy uczące się. *Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości. WNT, Warszawa*, 2008.

[77] Daniel Granato, Predrag Putnik, Danijela Bursać Kovačević, Jânio Sousa Santos, Verônica Calado, Ramon Silva Rocha, Adriano Gomes Da Cruz, Basil Jarvis, Oxana Ye Rodionova, and Alexey Pomerantsev. Trends in chemometrics: Food authentication, microbiology, and effects of processing. *Comprehensive Reviews in Food Science and Food Safety*, 17(3):663–677, 2018.

[78] Einar Etzold and Birgit Lichtenberg-Kraag. Determination of the botanical origin of honey by fourier-transformed infrared spectroscopy: an approach for routine analysis. *European Food Research and Technology*, 227(2):579–586, 2008.

[79] Eugeniusz Gatnar. Analiza dyskryminacyjna-stan aktualny i kierunki rozwoju. *Studia Ekonomiczne*, 152:42–58, 2013.

[80] C Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[81] Joseph G Bryan. The generalized discriminant function: mathematical foundation and computational routine. *Harvard Educational Review*, 21(2):90–95, 1951.

[82] Witold Malina. On an extended fisher criterion for feature selection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):611–614, 1981.

[83] Agata Kolakowska and Witold Malina. Fisher sequential classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(5):988–998, 2005.

[84] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[85] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[86] Michael D. Farrell and Russell M. Mersereau. On the impact of pca dimension reduction for hyperspectral detection of difficult targets. *IEEE Geoscience and Remote Sensing Letters*, 2(2):192–195, 2005.

[87] Yan Yang, Peng Cheng Nie, Wei Zhang, and Yong He. A novel method of pattern recognition for honey source base on Visible/Near infrared Spectroscopy: Genetic algorithm combined with support vector machine. *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010*, 1:519–523, 2010.

[88] S. Nawar and A. M. Mouazen. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. *Soil and Tillage Research*, 190(February):120–127, 2019.

[89] Mirosława Sztemberg-Lewandowska. Problemy decyzyjne w funkcjonalnej analizie głównych składowych. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 385:267–275, 2015.

[90] Svante Wold, Harold Martens, and Herman Wold. The multivariate calibration problem in chemistry solved by the pls method. In *Matrix pencils*, pages 286–293. Springer, 1983.

[91] Herman Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142, 1975.

[92] H Martens and T Naes. Methods for calibration. *Multivariate calibration*, 1:73–232, 1989.

[93] Nicolas Abdel-Nour. Chicken egg quality assessment from visible/near infrared observations. *ProQuest Dissertations and Theses*, (October):1–94, 2008.

[94] Eva M. Achata, Elena S. Inguglia, Carlos A. Esquerre, Brijesh K. Tiwari, and Colm P. O'Donnell. Evaluation of vis-nir hyperspectral imaging as a process analytical tool to classify brined pork samples and predict brining salt concentration. *Journal of Food Engineering*, 246(October 2018):134–140, 2019.

[95] K Michalowska, E Glowienka, et al. Multi-temporal data integration for the changeability detection of the unique słowinski national park landscape. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37:1017–1020, 2008.

[96] Wei Zeng, Ping Wang, Huiszhen Zhang, and Shenyang Tong. Qualitative and quantitative analyses of synthetic pigments in foods by using the branch and bound algorithm. *Analytica chimica acta*, 284(2):445–451, 1993.

[97] Mingchun Liu and Chunru Wan. Feature selection for automatic classification of musical instrument sounds. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 247–248. ACM, 2001.

[98] Sabrina Bouatmane, Mohamed Ali Roula, Ahmed Bouridane, and Somaya Al-Maadeed. Round-robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery. *Machine Vision and Applications*, 22(5):865–878, 2011.

[99] Iffat A Gheyas and Leslie S Smith. Feature subset selection in large dimensionality domains. *Pattern recognition*, 43(1):5–13, 2010.

[100] Katarzyna Stąpor. *Automatyczna klasyfikacja obiektów*. Akademicka Oficyna Wydawnicza EXIT, 2005.

[101] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern recognition and neural networks. In *Advanced Course on Artificial Intelligence*, pages 169–195. Springer, 1999.

[102] G Sahoo and Yugal Kumar. Analysis of parametric & non parametric classifiers for classification technique using weka. *International Journal of Information Technology and Computer Science (IJITCS)*, 4(7):43, 2012.

[103] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. adaptive computation and machine learning. *MIT Press*, 31:32, 2012.

[104] Aniruddha Ghosh, Fabian Ewald Fassnacht, PK Joshi, and Barbara Koch. A framework for mapping tree species combining hyperspectral and lidar data: Role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation*, 26:49–63, 2014.

[105] Katarzyna Stąpor, Paweł Błaszczyk, and Adrian Brückner. A comparative review of the selection methods for discovering differentially expressed genes in microarray experiments for classification. *Studia Informatica*, 27(4):37–52, 2006.

[106] Minjin Kim, Young-Hak Lee, and Chonghun Han. Real-time classification of petroleum products using near-infrared spectra. *Computers & Chemical Engineering*, 24(2-7):513–517, 2000.

[107] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[108] Henri A Vrooman, Chris A Cocosco, Fedde van der Lijn, Rik Stokking, M Arfan Ikram, Meike W Vernooij, Monique MB Breteler, and Wiro J Niessen. Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. *Neuroimage*, 37(1):71–81, 2007.

[109] Kunshan Huang, Shutao Li, Xudong Kang, and Leyuan Fang. Spectral–spatial hyperspectral image classification based on knn. *Sensing and Imaging*, 17(1):1, 2016.

[110] Eugeniusz Gatnar. *Nieparametryczna metoda dyskryminacji i regresji*. Wydaw. Naukowe PWN, 2001.

[111] Gordon V Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127, 1980.

[112] NL Hjort. *Pattern recognition and neural networks*. Cambridge university press, 1996.

[113] Inc StatSoft. Electronic statistics textbook. *Tulsa, OK: StatSoft*, page 34, 2013.

[114] Cecile Levasseur-Garcia, Sylviane Bailly, Didier Kleiber, and Jean-Denis Bailly. Assessing risk of fumonisin contamination in maize using near-infrared spectroscopy. *Journal of Chemistry*, 2015, 2015.

[115] Leo Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees. wadsworth & brooks. *Cole Statistics/Probability Series*, 1984.

[116] Wei-Yin Loh and Yu-Shan Shih. Split selection methods for classification trees. *Statistica sinica*, pages 815–840, 1997.

[117] Davide Bertelli, Maria Plessi, AG Sabatini, Massimo Lolli, and F Grillenzoni. Classification of italian honeys by mid-infrared diffuse reflectance spectroscopy (drifts). *Food Chemistry*, 101(4):1565–1570, 2007.

[118] Seng Khoon Teh, Wei Zheng, Khek Yu Ho, Ming Teh, Khay Guan Yeoh, and Zhiwei Huang. Diagnosis of gastric cancer using near-infrared raman spectroscopy and classification and regression tree techniques. *Journal of biomedical optics*, 13(3):034013, 2008.

[119] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[120] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[121] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[122] Philip Agre and Philip E Agre. *Computation and human experience*. Cambridge University Press, 1997.

[123] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[124] Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

[125] Sasan Karamizadeh, Shahidan M Abdullah, Mehran Halimi, Jafar Shayan, and Mohammad javad Rajabi. Advantage and drawback of support vector machine functionality. In *2014 international conference on computer, communications, and control technology (I4CT)*, pages 63–65. IEEE, 2014.

[126] Olivier Devos, Cyril Ruckebusch, Alexandra Durand, Ludovic Duponchel, and Jean-Pierre Huvenne. Support vector machines (svm) in near infrared (nir) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems*, 96(1):27–33, 2009.

[127] Bradley Efron and Robert J Tibshirani. *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*. Division of Biostatistics, Stanford University, 1995.

[128] Lukasz Krol. Distributed monte carlo feature selection: Extracting informative features out of multidimensional problems with linear speedup. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, pages 463–474. Springer, 2015.

[129] J Sunil Rao and Robert Tibshirani. The out-of-bootstrap method for model averaging and selection. *University of Toronto*, 1997.

[130] Rafał Adamczak. Zastosowanie sieci neuronowych do klasyfikacji danych doświadczalnych. *UMK, Toruń*, pages 1–235, 2001.

[131] Alain Zuur, Elena N Ieno, and Graham M Smith. *Analyzing ecological data*. Springer, 2007.

[132] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[133] M Dorozhovets and Zygmunt Lech Warsza. Propozycje rozszerzenia metod wyznaczania niepewności wyniku pomiarów wg przewodnika gum (1) uwzględnianie wpływu autokorelacji i nieadekwatności rozkładu wyników obserwacji w niepewności typu a. *Pomiary Automatyka Robotyka*, 11(1):6–15, 2007.

[134] S.R.S. Manual: Sr620 universal time interval counter. *Stanford Research Systems*, rev. 2.5 (4/2004), 1989.

[135] Wiesław Chmielnicki. Efektywne metody selekcji cech i rozwiązywania problemu wieloklasowego w nadzorowanej klasyfikacji danych. *Instytut Podstawowych Problemów Techniki Polskiej Akademii Nauk*, 2012.

[136] Alvin C Rencher. *Multivariate statistical inference and applications*. Wiley New York, 1998.

[137] Agnar Höskuldsson. Pls regression methods. *Journal of chemometrics*, 2(3):211–228, 1988.

[138] Zofia Lorenc, Sławomir Paśko, Olimpia Kursa, Anna Pakuła, and Leszek Sałbut. Spectral technique for detection of changes in eggshells caused by mycoplasma synoviae. *Poultry science*, 98(9):3481–3487, 2019.

[139] Zofia Lorenc, Sławomir Paśko, Anna Pakuła, Olimpia Kursa, and Leszek Sałbut. Spectral vis measurements for detection changes caused by of mycoplasma synoviae in flock of poultry. In *International Conference Mechatronics*, pages 422–429. Springer, 2019.

[140] Zofia Lorenc, Marek Karwowski, Sławomir Paśko, Olimpia Kursa, Anna Pakuła, and Leszek Sałbut. Combined optical coherence tomography and spectral technique for detection of changes in eggshells caused by mycoplasma synoviae. In *Speckle 2018: VII International Conference on Speckle Metrology*, volume 10834, page 108341M. International Society for Optics and Photonics, 2018.

[141] DS Zhu, JZ Pan, and Yong He. Identification methods of crop and weeds based on vis/nir spectroscopy and rbf-nn model. *Guang pu xue yu guang pu fen xi= Guang pu*, 28(5):1102–1106, 2008.

[142] Quansheng Chen, Jiewen Zhao, CH Fang, and Dongmei Wang. Feasibility study on identification of green, black and oolong teas using near-infrared reflectance spectroscopy based on support vector machine (svm). *Spectrochimica Acta Part A: Molecular*

and Biomolecular Spectroscopy, 66(3):568–574, 2007.

[143] Shih-Yu Chen, Yen Chieh Ouyang, and Chein-I Chang. Weighted radial basis function kernels-based support vector machines for multispectral image classification. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 4339–4342. IEEE, 2012.

[144] Patrícia Amaral Souza Tette, Letícia Rocha Guidi, Maria Beatriz de Abreu Glória, and Christian Fernandes. Pesticides in honey: A review on chromatographic analytical methods. *Talanta*, 149:124–141, 2016.

[145] Knut Faegri, Peter Emil Kaland, Knut Krzywinski, et al. *Textbook of pollen analysis*. Number Ed. 4. John Wiley & Sons Ltd., 1989.

[146] Sahameh Shafiee, Gerrit Polder, Saeid Minaei, Nasrolah Moghadam-Charkari, Saskia Van Ruth, and Piotr M Kuś. Detection of honey adulteration using hyperspectral imaging. *IFAC-PapersOnLine*, 49(16):311–314, 2016.

[147] Yan Yang, Peng-Cheng Nie, Wei Zhang, and Yong He. A novel method of pattern recognition for honey source based on visible/near infrared spectroscopy: Genetic algorithm combined with support vector machine. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, pages 519–523. IEEE, 2010.

[148] Zofia Lorenc, Sławomir Tomczewski, and Leszek Sałbut. Graphene nanoplatelets size analysis based on sample transparency. *Photonics Letters of Poland*, 7(4):118–120, 2015.