WOJSKOWA AKADEMIA TECHNICZNA
im. Jarosława Dąbrowskiego
WYDZIAŁ MECHATRONIKI I LOTNICTWA

PRACA DYPLOMOWA

STUDIA WYŻSZE

sierż. pchor. inż. Dawid BREJECKI
(stopień, imiona i nazwisko studenta)

Metody statystyczne w klasyfikacji pocisków na podstawie danych pochodzących ze strzelań

Statistical methods in the classification of projectiles based on data from the shooting range
(temat pracy dyplomowej w języku polskim i języku angielskim)

Mechatronika, Eksploatacja przeciwlotniczych zestawów rakietowych
(kierunek i specjalność studiów)

ppłk dr inż. Dariusz RODZIK
(stopień wojskowy, tytuł/stopień naukowy, imię i nazwisko promotora pracy dyplomowej)

Warszawa – 2020

(strona celowo zostawiona pusta)

Zadania do pracy dyplomowej

Zadania do pracy dyplomowej

Spis treści

(strona celowo zostawiona pusta)

Wykaz oznaczeń

a	liczba klas (grup)
A	amplituda zaburzenia ciśnienia
b_0	stała w funkcji dyskryminacyjnej LDA
b_p	współczynnik dyskryminacyjny Fischera
c	prędkość propagacji dźwięku w powietrzu
C	parametr kosztu kary w SVM
czas	zmienna w modelu określająca czas trwania fali N w [ms]
d	odległość czujnika pomiarowego od trajektorii lotu pocisku
f	funkcja gęstości prawdopodobieństwa
$\boldsymbol{g}$	macierz wewnątrzgrupowej sumy kwadratów
$\boldsymbol{h}$	macierz międzygrupowej sumy kwadratów
k, c	numery klas (grup)
K	liczba punktów odniesienia w metodzie najbliższych sąsiadów
kaliber	zmienna w modelu określająca kaliber pocisku w [mm]
KNN	metoda K-najbliższych sąsiadów (ang. *K-nearest neighbours*)
l	długość pocisku
L	funkcja wiarygodności
L_{UR}	wartość funkcji wiarygodności dla pełnego modelu
L_R	wartość funkcji wiarygodności dla modelu zawierającego tylko wyraz wolny
LDA	liniowa analiza dyskryminacyjna (ang. *linear discriminant analysis*)
LR	wartość testu ilorazu wiarygodności
M	liczba Macha
M6	oznaczenie mikrofonu
MB	statystyka testu M-Boxa
n	liczba obserwacji
n_k	liczba obserwacji w danej klasie
odl	zmienna w modelu określająca odległość czujników od trajektorii lotu pocisku w [m]

QDA kwadratowa analiza dyskryminacyjna (ang. *quadratric discriminant analysis*)

p liczba predyktorów

p – value, p – najniższy poziom istotności, przy którym odrzuca się hipotezę zerową

P_0 ciśnienie powietrza

R_k^2 współczynnik korelacji wielorakiej

SE błąd standardowy

SVM metoda wektorów nośnych (ang. *support vector machine*)

T czas trwania fali N

T_k współczynnik tolerancji

V prędkość pocisku

W wartość statystyki Shapiro-Wilka

WAT Wojskowa Akademia Techniczna

WITU Wojskowy Instytut Techniczny Uzbrojenia

$\boldsymbol{x}$ wektor wartości obserwacji

$\boldsymbol{X}$ wektor predyktorów

$\boldsymbol{Y}$ wektor klas (grup)

Z wartość funkcji dyskryminacyjnej w LDA

$\boldsymbol{\delta_k}$ wektorowa postać funkcji dyskryminacyjnej w LDA i QDA

ζ_1, ζ_2 współczynniki Whithama

θ_M kąt rozwarcia stożka

λ wartość statystyki Lambda-Wilksa

λ_k^{cz} cząstkowy współczynnik Lambda-Wilksa

$\boldsymbol{\mu_k}$ wektor wartości średnich zmiennych w danej klasie

$\widehat{\boldsymbol{\pi}}_{\boldsymbol{k}}$ wektor częstości w LDA i QDA

$\sum$ macierz kowariancji połączonych

$\sum_k$ macierz kowariancji wewnątrz grupy

$\sum_s$ międzygrupowa macierz kowariancji

τ_i czas przybycia fali uderzeniowej do czujników pomiarowych

ϕ średnica pocisku

Wstęp

W warunkach współczesnego pola walki, kluczowym czynnikiem przesądzającym o zwycięstwie nie jest już przewaga liczebna, czy nawet umiejętności dowódców, lecz informacja. Informacja o rodzaju używanej broni i amunicji przez przeciwnika może być bardzo istotna na polu walki. We współczesnych systemach identyfikacji pocisków przeważnie stosowane są techniki radarowe, ale konsekwentnie rozwijane jest także podejście akustyczne. Ten rodzaj identyfikacji opiera się na detekcji unikalnych właściwości akustycznej fali uderzeniowej (nazywanej falą N), generowanej przez pociski poruszające się z prędkością naddźwiękową. Główną zaletą akustycznej lokacji jest niski koszt implementacji oraz uniwersalność tej metody [10].

Literatura przedmiotu liczy niewiele publikacji związanych z identyfikacją pocisków na podstawie parametrów fali N. Główne wysiłki badaczy są skierowane na zrozumienie zjawiska powstawania i propagacji zaburzenia w ośrodku oraz wykorzystanie fali N do wyznaczenia położenia strzelca [6]. W 1995 roku, na polecenie armii, amerykańscy naukowcy skonstruowali modele klasyfikujące pociski na podstawie parametrów fali N stosując liniową analizę dyskryminacyjną [10]. W 2007 roku powstał kompletny system identyfikacji położenia strzelca, który także rozróżniał kaliber pocisku, jednakże wykorzystywał do tego teoretyczny wzór Whithama, co ma swoje wady [2]. W 2011 roku chińscy naukowcy zaproponowali identyfikację pocisku z użyciem metody wektorów nośnych [7]. W literaturze wciąż brak jest jednak badań skupiających się na wskazaniu najskuteczniejszego klasyfikatora rozróżniającego kaliber pocisku na podstawie parametrów charakterystycznych fali N.

Celem pracy jest analiza porównawcza sześciu najczęściej wykorzystywanych do klasyfikacji metod statystycznych i wybór najskuteczniejszego klasyfikatora rozróżniającego pociski na podstawie parametrów charakterystycznych fali N. Opracowane klasyfikatory stanowią nową wiedzę w tym temacie.

Z uwagi na dostępny zakres danych ze strzelań, opracowane modele identyfikują pociski kalibru 5,56, 7,62, 23 oraz 35 [mm]. W literaturze przedmiotu nie ma obecnie opracowanych modeli klasyfikujących pociski o podanych kalibrach. Zaletą proponowanego opracowania jest oparcie zbudowanych modeli klasyfikacyjnych na danych ze strzelań w różnych warunkach środowiskowych. W pracy zebrano i uporządkowano dane ze

wszystkich badań przeprowadzonych na Wydziale Mechatroniki WAT, które mogły być wykorzystane w opracowaniu modeli identyfikujących pociski.

Przeanalizowane zostały tradycyjne metody klasyfikacji tj. liniowa i kwadratowa analiza dyskryminacyjna oraz regresja logistyczna, a także nowoczesne metody uczenia maszynowego: K-najbliższych sąsiadów, wektorów nośnych oraz naiwny klasyfikator Bayesa.

W rozdziale pierwszym przedstawiona jest teoria na temat zjawiska powstawania fali N, szczególnie pod kątem wykorzystania jej parametrów w klasyfikacji pocisków, a także aktualny stan wiedzy na temat wykorzystania metod statystycznych do klasyfikacji pocisków na podstawie parametrów fali N. W rozdziale drugim opisane są wybrane metody statystyczne oraz sposób ich wykorzystania przy wyznaczaniu modeli klasyfikacyjnych. Proces pozyskania i przygotowania zbioru danych ze strzelań opisany został w rozdziale trzecim, natomiast w rozdziale czwartym – empirycznym, przedstawiono wyniki przeprowadzonych badań i analiz porównawczych, a także sformułowano wnioski końcowe.

Wszystkie algorytmy klasyfikacji, zamieszczone w pracy, zostały zaimplementowane w języku programowania R z użyciem programu RStudio. Do wyznaczenia parametrów fali N na podstawie wyeksportowanych plików pomiarowych użyto programu napisanego przez autora w języku Python oraz programów MS Excel i PicoScope.

1. Podstawy teoretyczne

1.1. Analiza stanu wiedzy

We współcześnie wykorzystywanych systemach identyfikacji pocisków (szczególnie broni palnej), dane o obiekcie są najczęściej zbierane z pomocą technik radiolokacyjnych. Według najnowszych trendów dane te są następnie wykorzystywane do klasyfikacji za pomocą ukrytych modeli Markova lub sieci neuronowych [3÷5].

Jednakże konsekwentnie rozwijane jest także podejście bazujące na technologii lokacji akustycznej. Techniki akustyczne są przede wszystkim tańsze i łatwiejsze w implementacji od technik radiolokacyjnych. Ponadto cechują się dużą uniwersalnością - mogą być stosowane w mobilnych systemach identyfikacji [6], [10].

W literaturze przedmiotu niewiele jest publikacji na temat zastosowania technik akustycznej lokacji obiektów do klasyfikacji pocisków. Pierwszą pozycją w literaturze informującą na temat możliwości klasyfikacji pocisków na postawie znajomości sygnatury akustycznej jest raport techniczny amerykańskich naukowców [10]. W niniejszym projekcie zbudowano modele oparte na znajomości wartości parametrów fali N, a także odległości czujników do trajektorii lotu oraz prędkości pocisku. Do rozróżnienia kalibrów wykorzystano tylko metodę liniowej analizy dyskryminacyjnej.

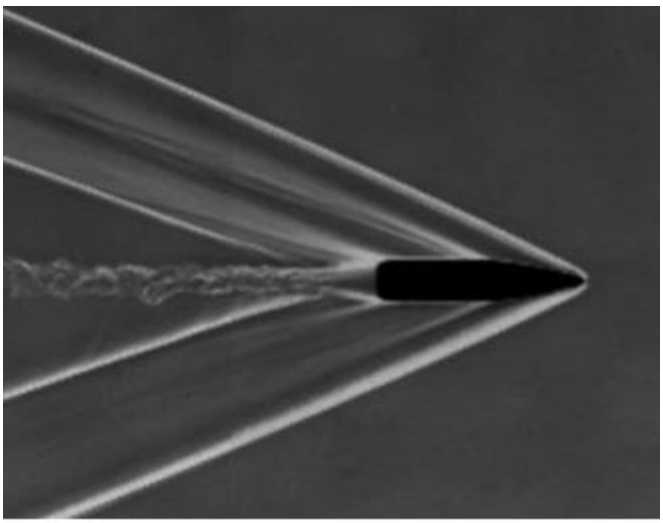
W roku 2007, w kompletnym opracowaniu [2] amerykańscy naukowcy wykonali mobilny system natychmiastowego rozpoznawania położenia strzelca i trajektorii lotu pocisku, bazujący na rejestracji akustycznej. System został wprowadzony na rynek i jest wykorzystywany przez żołnierzy armii USA. Określenie kalibru pocisku nie było jednak nadrzędnym celem działania tego systemu. System wyznaczał kaliber nie na podstawie klasyfikatora, tylko poprzez podstawienie zarejestrowanych danych do modelu Whithama. W przypadku braku rejestracji prędkości pocisku, kaliber nie mógł być wyznaczony, ponadto to podejście nie uwzględniało zmiany wartości parametrów fali N w różnych warunkach środowiskowych.

W 2011 roku chińscy badacze uznali podejście akustyczne jako efektywne w klasyfikacji pocisków i zaproponowali metodę wektorów nośnych (SVM) do rozróżniania kalibrów na podstawie parametrów fali N [7].

Niniejsza praca jest prawdopodobnie pierwszą, w której przeanalizowano sześć najczęściej wykorzystywanych metod statystycznych mogących być użytecznymi w klasyfikacji pocisków na postawie parametrów fali N pozyskiwanych z pomocą czujników akustycznych.

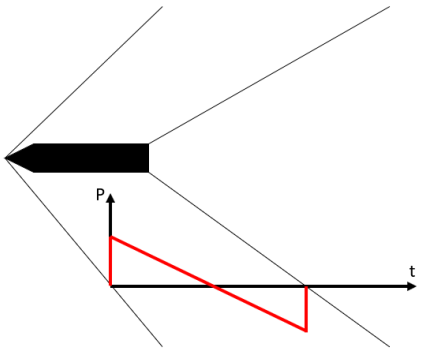
1.2. Fala N

Pocisk lub dowolne inne wystrzelone ciało osiowosymetryczne, poruszające się w powietrzu z prędkością naddźwiękową, generuje akustyczną falę uderzeniową (ang. *shock wave*). Podczas lotu pocisku, powietrze jest sprężane naprzeciw jego czoła. Ponieważ obiekt porusza się szybciej niż czas potrzebny powietrzu do rozprężenia się przed obiektem, generowane jest stożkowe zaburzenie ośrodka, którego front powoduje gwałtowny wzrost ciśnienia [10], jak to jest pokazane na rys. 1.



Rysunek 1. Pocisk z falą uderzeniową [35].

Sprężone powietrze rozszerza się i tym samym ciśnienie lokalne zmniejsza się znacznie poniżej wartości początkowej. Na końcu przejście dolnej części fali uderzeniowej powoduje kolejny skokowy wzrost ciśnienia. Jeżeli czujnik ciśnienia umieszczony jest na drodze propagacji fal uderzeniowych, to zarejestruje zaburzenie ciśnienia o charakterystycznym kształcie, tzw. falę N, co jest zobrazowane na rys. 2.



Rysunek 2. Powstawanie fali N.

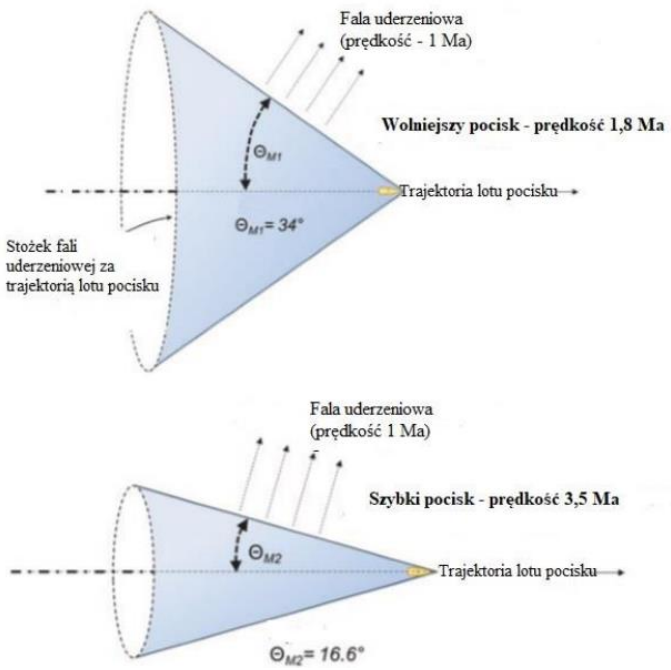
Kształt powstającego zaburzenia przybiera charakterystyczną stożkową postać, zwaną stożkiem Macha. Im kąt rozwarcia stożka jest mniejszy, tym krótszy jest czas trwania powstałego zaburzenia ciśnienia. Kąt rozwarcia stożka zaburzenia można opisać wzorem [36]:

$$\theta_M = \arcsin\left(\frac{1}{M}\right)$$

gdzie M – liczba Macha, która jest stosunkiem prędkości pocisku V do prędkości propagacji dźwięku c:

$$M = \frac{V}{c}.$$

Wynika stąd, że im wyższa prędkość pocisku V, tym mniej rozwarty jest stożek fali uderzeniowej (rys. 3):



Rysunek 3. Kształt stożka zaburzenia w zależności od prędkości pocisku [37].

Z analizy literatury wynika, że prędkość pocisku nie jest jedynym parametrem wpływającym na przebieg fali N. Model matematyczny fali N dla ciał osiowosymetrycznych (w tym pocisków) szczegółowo opisał G. B. Whitham w publikacjach [38], [39]. Postulaty Whithama są do dziś podstawą obliczeń parametrów fali uderzeniowej. Zgodnie z nimi amplitudę A oraz czas trwania T fali N można przedstawić jako funkcję średnicy ϕ, długości l, i prędkości pocisku V oraz odległości d między pozycją czujnika pomiarowego, a trajektorią lotu pocisku.

Wspomnianą zależność opisują następujące wzory:

$$T = \frac{1{,}82 \cdot M \cdot \phi \cdot d^{\zeta_1}}{l^{1/4} \cdot c \cdot (M^2-1)^{3/8}}\ [s], \qquad (*)$$

$$A = \frac{0{,}53 \cdot P_0 \cdot (M^2-1)^{1/8} \cdot \phi}{d^{\xi_2} \cdot l^{1/4}}\ [Pa\ ],$$

gdzie P_0 – ciśnienie powietrza, ζ_1, ζ_2 – współczynniki Whithama wyznaczane w sposób doświadczalny.

Wyróżnia się ponadto czas przybycia fali uderzeniowej do czujników pomiarowych τ_i :

$$\tau_i = \frac{s_i}{c},$$

gdzie s_i – odległość w linii prostej między źródłem zaburzenia, a czujnikiem pomiarowym w danym momencie.

Późniejsze badania naukowców [8],[40] wykazały, że teoretyczne wartości parametrów fali N obliczone z modelu Whithama są szczególnie zgodne z uzyskanymi wartościami podczas eksperymentów dla pocisków o dość regularnym kształcie i liczbie Macha poniżej 3. Dla wyższej liczby Macha, wartości teoretyczne parametrów fali N coraz bardziej odbiegają od eksperymentalnych. Empirycznie wyznaczony czas trwania fali N jest większy od teoretycznego.

W związku z tym, że falę N można opisać modelem matematycznym Whithama, daje to podstawy do postawienia hipotezy, że:

> *Można klasyfikować pociski na podstawie analizy wartości pozyskanych parametrów charakterystycznych fali N.*

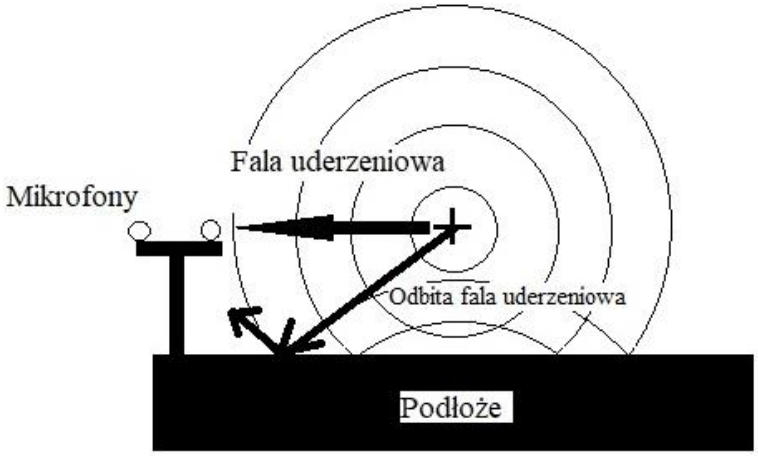
Zgodnie z postulatami Whithama, oprócz wymiarów pocisku, na przebieg fali N mają wpływ także: odległość czujników od trajektorii lotu pocisku oraz liczba Macha, czyli stosunek

prędkości pocisku do prędkości dźwięku w danych warunkach. Badania [9] wykazały, że temperatura otoczenia nie ma istotnego wpływu na parametry fali N w typowym przedziale temperatur [± 50ºC].

Tak jak czas trwania fali N jest łatwo mierzalną wartością (wystarczy zastosować odpowiednio dokładny licznik czasu), tak mierzenie amplitudy fali N nie jest już takie proste. Każdy czujnik ciśnienia w zależności od jego czułości własnej, inaczej reaguje na zmianę ciśnienia otoczenia (posiada inną czułość [V/Pa]). Powstaje więc problem ciągłej kalibracji torów pomiarowych, co jest mało praktyczne w świetle dalszych zastosowań. Ponadto pomiary amplitudy ciśnienia cechują się dość dużym błędem. Z tych powodów dane związane z amplitudą ciśnienia nie są w pełni miarodajną zmienną statystyczną. Do podobnych wniosków doszli naukowcy w pracy [10].

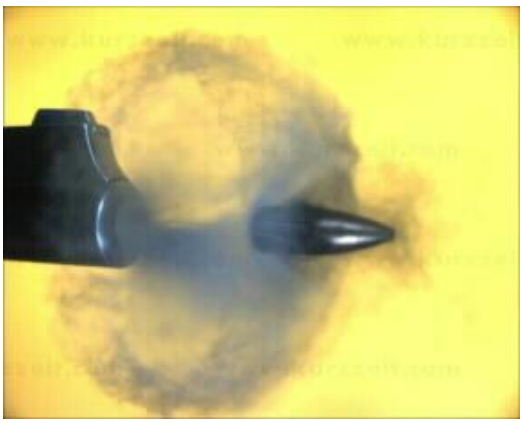
W świetle powyższych teoretycznych rozważań, model klasyfikujący pociski powinien zawierać następujące zmienne: czas trwania fali N, prędkość pocisku oraz odległość trajektorii od czujników pomiarowych.

Jednakże zarejestrowanie przebiegu fali N identycznego z teoretycznym modelem jest bardzo trudne, a w większości warunków, niemożliwe. Fala uderzeniowa, podobnie jak inne fale fizyczne, podlega odbiciu, tłumieniu, absorpcji, dyfrakcji i innym modyfikacjom podczas jej propagacji. Mikrofon lub inny czujnik akustyczny zarejestruje fale docierające ze źródła, jak i fale docierające później z innych kierunków z powodu odbić i rozproszenia. Ponadto istotny wpływ mają także czynniki atmosferyczne. Największym problemem podczas tworzenia modelu klasyfikującego pociski, wskazanym przez naukowców w pracy [10], było występowanie fali odbitej, zakłócającej idealny przebieg fali N. Powstawanie odbitej fali N jest schematycznie przedstawione na rys. 4:



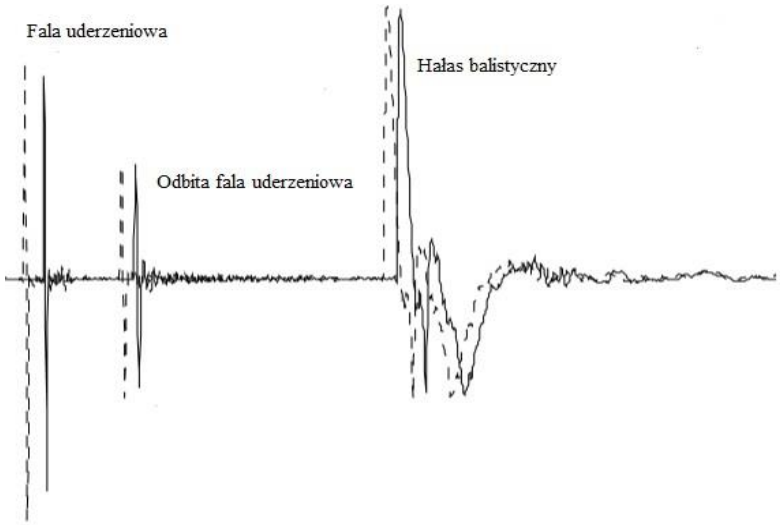
Rysunek 4. Powstawanie fali odbitej.

W typowym pomiarze strzału rejestruje się zarówno falę N, jak i wspomnianą wcześniej falę odbitą. Ponadto występuje tzw. hałas balistyczny, czyli dźwięk pochodzący od gazów prochowych wydostających się z lufy. Hałas balistyczny jest charakterystycznym dźwiękiem słyszalnym podczas oddawania strzału i nie może być mylony z falą uderzeniową od pocisku. Tak jak zredukowanie hałasu balistycznego jest możliwe poprzez zastosowanie tłumika, tak nie jest to możliwe w przypadku fali uderzeniowej od pocisku. Zobrazowanie zaburzenia ośrodka wywołanego hałasem balistycznym jest przedstawione na rys. 5.



Rysunek 5. Zaburzenie ośrodka spowodowane gazami wychodzącymi z lufy [11].

Na rys. 6 przedstawiono typowy przebieg rejestracji strzału:

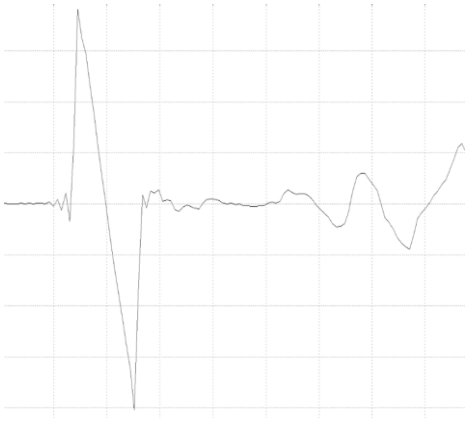


Rysunek 6. Przebieg rejestracji strzału [12].

W typowym przebiegu rejestracji strzału (rys. 6) pierwszym zarejestrowanym zaburzeniem ciśnienia jest fala N. Następnie w różnej kolejności pojawiają się: hałas balistyczny oraz odbita fala uderzeniowa.

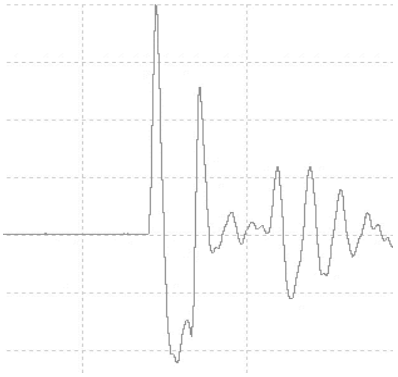
Odbita fala uderzeniowa przyjmuje kształt N tak jak macierzysta fala, oraz charakteryzuje się zmniejszoną amplitudą. Hałas balistyczny posiada nieregularny kształt, ale można wyróżnić w nim charakterystyczny skok ciśnienia. Często posiada wyższą amplitudę niż fala uderzeniowa od pocisku. Ma to istotne znaczenie podczas analizy zarejestrowanych przebiegów. Nie można z góry przyjmować, że fala o największej wartości szczytowej jest falą N, chyba że wiadomo, iż hałas balistyczny nie został zarejestrowany.

Przykład niemal idealnie zarejestrowanej fali N jest przedstawiony na rys. 7:



Rysunek 7. Poprawnie zarejestrowana fala N [31].

Z zarejestrowanego przebiegu na rys. 7 wynika, że fala odbita niemalże nie występuje i co najważniejsze nie „nachodzi" na falę N, nie zakłócając jej przebiegu. Hałas balistyczny nie został zarejestrowany. Zakłócony przebieg fali N jest przedstawiony na rys. 8:



Rysunek 8. Fala N z zachodzącą na nią falą odbitą [34].

Z przebiegu na rys. 8 wynika, że fala odbita dotarła do czujników pomiarowych jeszcze przed końcem rejestracji fali N i spowodowała nieregularność jej przebiegu. Pojawiła się ona w chwili rejestracji drugiego skoku ciśnienia fali uderzeniowej, powodując zwiększenie czasu trwania fazy ujemnej. Tym samym czas trwania fali N został fałszywie wydłużony.

Fala odbita przedstawiona na rys. 8 osiągnęła dużą amplitudę, znacznie przekraczając połowę amplitudy fali macierzystej. Świadczy to o tym, że fala N odbiła się od twardej powierzchni. Ponadto powierzchnia odbijająca znajdowała się blisko czujników akustycznych, co spowodowało „nałożenie się" fali odbitej na falę N. Zredukowanie hałasu balistycznego podczas pomiarów eksperymentalnych jest możliwe poprzez umieszczenie czujników pomiarowych odpowiednio daleko od lufy. Spowoduje to, że energia hałasu balistycznego zmaleje do nieistotnych wartości jeszcze przed jego dotarciem do mikrofonów pomiarowych.

Z analizy literatury wynika, że na przebieg fali N mają wpływ również inne czynniki środowiskowe. Propagacja fali dźwiękowej może znacznie różnić się ze względu na warunki atmosferyczne, dyfrakcję wokół przeszkadzających obiektów, odbicia od ziemi i innych powierzchni. Co więcej, na propagację dźwięku może wpływać wiatr, zmiany temperatury oraz pochłanianie atmosferyczne zależne od częstotliwości. Zwykłe, słyszalne dźwięki mieszczą się w zakresie ciśnienia, który jest dobrze modelowany liniowymi równaniami różniczkowymi fal, ale fale uderzeniowe spowodowane naddźwiękowym ruchem pocisków zachowują się w powietrzu nieliniowo. Możliwe jest przewidywanie właściwości akustycznych danego środowiska przy użyciu standardowych metod matematycznych, jednak właściwości absorpcji, tłumienia i odbicia od obiektów muszą być znane dla liniowych i nieliniowych propagacji. Typowe dźwięki obejmują zakres częstotliwości od poniżej 10 Hz do ponad 40 kHz. Długości fali napotykane w powietrzu mogą wahać się od ponad 30 m przy niskich częstotliwościach, do mniej niż 0,01 m dla najwyższych częstotliwości. Zakres długości fal jest duży, oznacza to, że właściwości dyfrakcyjne i absorpcyjne będą znacznie się różnić w zależności od spektrum źródła dźwięku [12].

Opady atmosferyczne zwykle nie mają znaczącego wpływu na fale dźwiękowe, jednak mają one wpływ na wilgotność, wiatr i zmiany temperatury, a także generują hałas akustyczny.
Duży wpływ na warunki pomiaru ma wiatr. Fala dźwiękowa jest przenoszona w ruchomej masie powietrza. Poprzez działanie wiatru ruch czoła fali uderzeniowej będzie składał się z sumy wektorowej sferycznie rozwijającego się wektora dźwięku oraz wektora wiatru. Efekt wiatru można postrzegać jako przesunięcie punktu początkowego propagacji dźwięku. Czoło fali jest systematycznie przesuwane przez wiatr tak, że po dotarciu fali do mikrofonów, przesunięte zostaje pozorne położenie źródła dźwięku i trajektorii pocisku. Prędkość wiatru jest zwykle szybsza w górnych partiach i mniejsza na dole. To powoduje, że fale dźwiękowe propagujące z wiatrem są wygięte w górę, a te które propagują się pod wiatr, są wygięte w dół. Chociaż prędkości wiatru są dużo niższe w porównaniu z prędkością dźwięku, mają

one wpływ, gdyż zmiana wiatru powoduje zależną od kierunku zmianę prędkości dźwięku i co za tym idzie, przesunięcie częstotliwości, podobne do efektu Dopplera [12].

Wpływ nierównomiernej temperatury może również być znaczący. Temperatura powietrza w atmosferze nie jest jednorodna. W dzień, szczególnie w miesiącach letnich powierzchnia ziemi jest cieplejsza niż powietrze. W tej sytuacji propagacja dźwięku ma tendencje do pochylania się nieco w górę ze względu na temperaturę. Czoło fali w ciepłym powietrzu w pobliżu powierzchni ziemi porusza się szybciej niż czoło fali w chłodniejszym powietrzu wyżej nad ziemią. Zimą lub w nocy, gdy temperatura powierzchni ziemi jest niższa, fale dźwiękowe są pochylane w dół. Połączone efekty wiatru i zmian temperatury mogą spowodować, że poziomy dźwięku mierzone w pewnej odległości od źródła będą bardzo różne od przewidywań opartych na rozkładzie geometrycznym i rozważaniach dotyczących absorpcji atmosferycznej. Różnice te mogą wynosić 20 dB lub więcej na odległościach kilkuset metrów [12].

Kolejnym czynnikiem jest powierzchnia ziemi i przeszkody. Odgłosy wystrzałów rozprzestrzeniających się nad ziemią ulegną osłabieniu poprzez straty energii akustycznej spowodowanej rozpraszaniem. Gładkie i twarde podłoże generalnie będzie mniej absorbowało niż szorstkie jak np. roślinność. Wyższe częstotliwości (krótsze fale) są tłumione bardziej niż niższe częstotliwości. Pomiary na obszarach zalesionych pokazują, że absorpcja i rozpraszanie mogą osiągnąć znaczne tłumienie. Problemem może również być tłumienie poprzez przeszkodę zasłaniającą linię widzenia między źródłem, a czujnikiem akustycznym [12].

Wilgotność względna powietrza powoduje pochłanianie dźwięku zależne od częstotliwości z powodu relaksacji termicznej cząsteczek. Stwierdzono, że tłumienie wzrasta monotonicznie wraz ze wzrostem częstotliwości i jest największe dla wilgotności względnej w zakresie 10-30% [12].

Znaczenie czynników atmosferycznych zależy więc od rozważanej sytuacji.

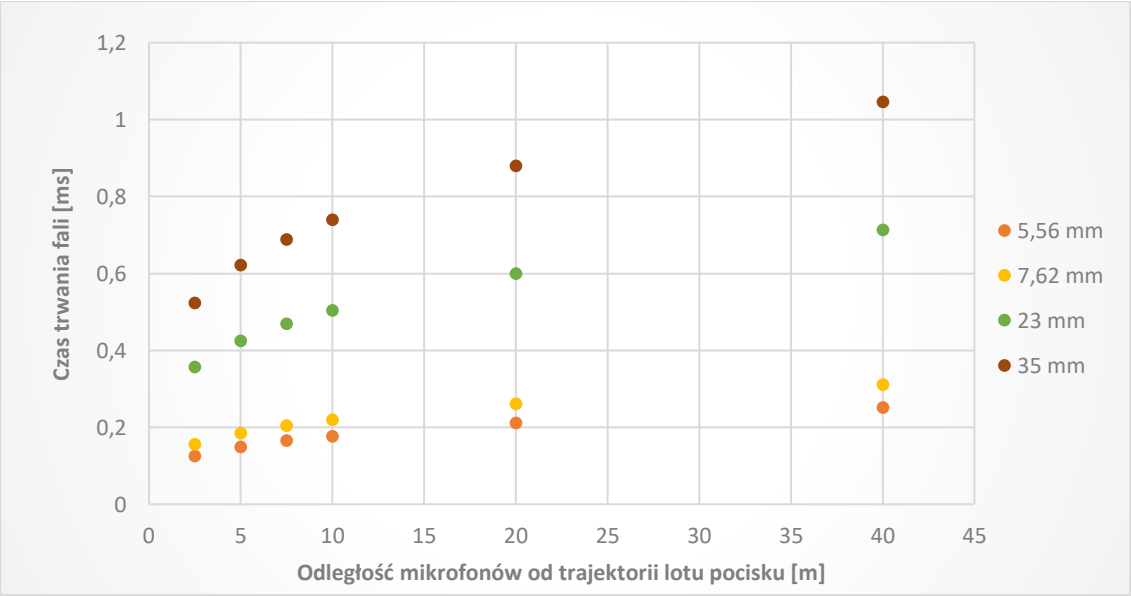
Zbudowanie modelu klasyfikującego pociski o największej wiarygodności wymaga zatem posiadania dużej ilości danych empirycznych dla różnych warunków środowiskowych.

W tabeli 1 przedstawione są teoretyczne czasy trwania fali N obliczone na podstawie wzoru (*) dla kalibrów 5,56, 7,62, 23 oraz 35 [mm].

Tabela 1. Wartości czasu trwania fali N wyznaczone z modelu Whithama.

Odległość czujników od trajektorii lotu [m]	Czas trwania fali N [ms] dla kalibru:			
	5,56 mm	7,62 mm	23 mm	35 mm
2,5	0,125	0,155	0,357	0,523
5	0,149	0,185	0,424	0,622
7,5	0,165	0,204	0,469	0,688
10	0,177	0,220	0,504	0,740
20	0,211	0,261	0,600	0,879
40	0,251	0,311	0,713	1,046

Dane z tabeli 1 są zaprezentowane na wykresie (rys. 9):



Rysunek 9. Wykres czasu trwania fali dla różnych kalibrów.

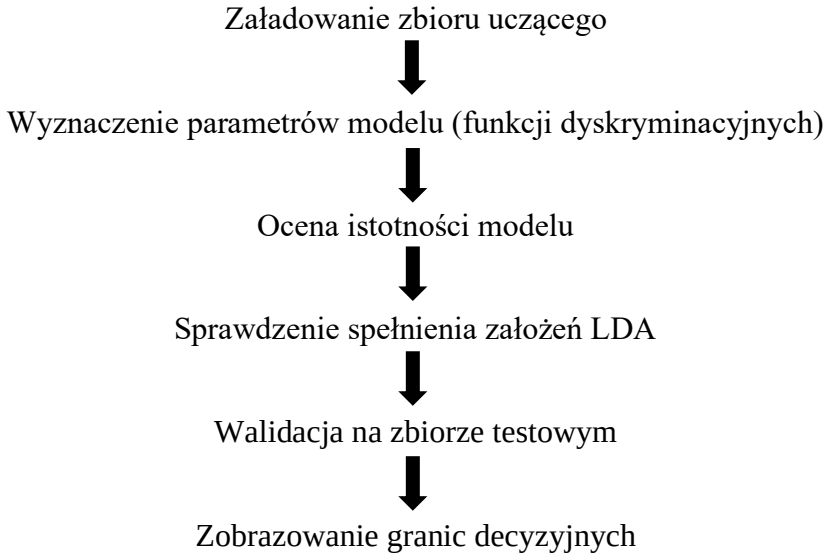
Wartości czasów trwania fali N poszczególnych kalibrów są na tyle różne, że *są przesłanką do możliwości stworzenia modelu klasyfikującego*, szczególnie dobrze rozróżniającego pociski kalibru 35 mm od 23 mm oraz 7,62 mm od 5,56 mm.

2. Analiza wykorzystanych metod statystycznych

W pracy zebrano wybrane metody klasyfikacji danych liczbowych: tradycyjne, dobrze poznane metody liniowej i kwadratowej analizy dyskryminacyjnej oraz regresji logistycznej, a także współczesne algorytmy uczenia maszynowego używane w klasyfikacji: K – najbliższych sąsiadów, wektorów nośnych oraz naiwny klasyfikator Bayesa.

Na początku kolejnych podrozdziałów przedstawiono schematy przyjętej procedury tworzenia klasyfikatora dla każdej z rozpatrywanych metody, a następnie opisano poszczególne etapy tej procedury.

2.1. Liniowa analiza dyskryminacyjna (LDA)



Wyznaczenie parametrów modelu

Liniowa analiza dyskryminacyjna (ang. *linear discriminant analysis*) jest wyprowadzona z prostych modeli probabilistycznych, które modelują warunkowy rozkład danych $P(\boldsymbol{X} = x|\boldsymbol{Y} = k)$ dla każdej klasy k. Prognozy można uzyskać stosując regułę Bayesa:

$$P(\boldsymbol{Y} = k|\boldsymbol{X} = x) = \frac{P(\boldsymbol{X} = x|\boldsymbol{Y} = k)P(\boldsymbol{Y} = k)}{P(\boldsymbol{X})} = \frac{P(\boldsymbol{X} = x|\boldsymbol{Y} = k)P(\boldsymbol{Y} = k)}{\sum_l P(\boldsymbol{X} = x|\boldsymbol{Y} = l) \cdot P(\boldsymbol{Y} = l)}$$

gdzie $\boldsymbol{X}$ oznacza wektor predyktorów, a $\boldsymbol{Y}$ wektor klas. Wybiera się klasę k, dla której prawdopodobieństwo warunkowe jest największe [43].

W LDA $P(\boldsymbol{X} = x|\boldsymbol{Y} = k)$ jest modelowane jako wielowymiarowy rozkład normalny z gęstością [43]:

$$P(\boldsymbol{X} = x|\boldsymbol{Y} = k) = \frac{1}{(2\pi)^{d/2}|\sum|^{1/2}} e^{-1/2(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\sum^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}$$

By użyć tego modelu jako klasyfikatora, należy wyznaczyć także prawdopodobieństwa a priori klas $P(\boldsymbol{Y} = k)$ [43]. W LDA wyznacza się to prawdopodobieństwo jako stosunek liczby obserwacji n_k w k-tej klasie do liczby wszystkich obserwacji n [42]:

$$P(\boldsymbol{Y} = k) = \frac{n_k}{n}$$

$\sum$ oznacza macierz kowariancji połączonych (ang. *pooled within-groups covariance matrix*) Estymatorem $\sum$ jest średnia ważona macierzy kowariancji z poszczególnych grup $\sum_k$ z wagami [45]:

$$\frac{n_k - 1}{n - p}$$

gdzie p oznacza liczbę predyktorów. W LDA zakłada się, że macierz kowariancji w każdej grupie jest równa.

Estymatorem $\boldsymbol{\mu}_k$ jest wartość średnia zmiennej dla obserwacji w danej klasie [45].

Niech $P(\boldsymbol{Y} = k) = \widehat{\boldsymbol{\pi}}_k$ i $P(\boldsymbol{Y} = k|\boldsymbol{X} = x) = \delta_k(x)$, wtedy obserwację $\boldsymbol{X} = x$ przypisuje się do klasy k, dla której funkcja:

$$\delta_k(x) = \boldsymbol{x}^T\sum^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k{}^T\sum^{-1}\boldsymbol{\mu}_k + \log(\widehat{\boldsymbol{\pi}}_k)$$

przyjmuje wartość maksymalną. Jest to postać tzw. dyskryminacyjnej funkcji liniowej Fischera [44]. Po podstawieniu do niej wartości obserwacji dla poszczególnych klas, wyznaczone zostają liniowe funkcje dyskryminacyjne dla każdej klasy w postaci [14]:

$$Z = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p.$$

Z – jest zmienną zależną (objaśnianą) – wartością funkcji dyskryminacyjnej,

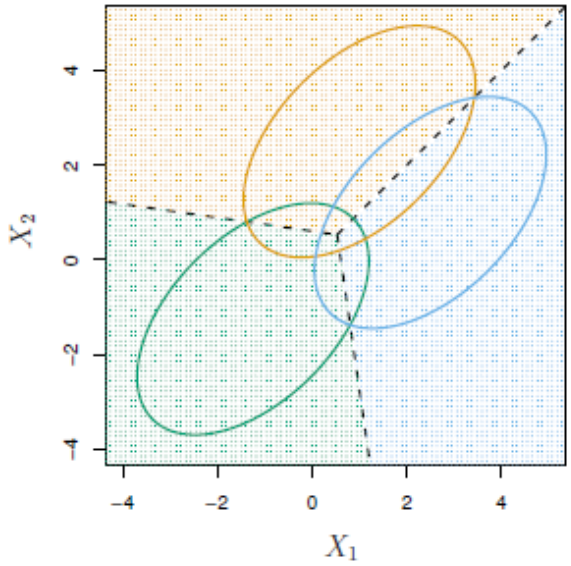
b_p – współczynniki dyskryminacyjne Fischera dla i = 1, 2,…, p,

b_0 – stała,

X_p – predyktory dla i =1, 2,… p.

Granice decyzyjne

Na rys. 10 przedstawione jest zobrazowanie granic decyzyjnych dla przykładowych danych.



Rysunek 10. Granice decyzyjne w LDA [44].

Granice decyzyjne na rys. 10 są zaznaczone liniami przerywanymi. Stanowią je zbiory punktów, dla których $\delta_k(x) = \delta_c(x)$, dla $c \neq k$, czyli [44]:

$$\boldsymbol{x}^T\sum^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k{}^T\sum^{-1}\boldsymbol{\mu}_k + \log(\hat{\boldsymbol{\pi}}_k) = \boldsymbol{x}^T\sum^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c{}^T\sum^{-1}\boldsymbol{\mu}_c + \log(\hat{\boldsymbol{\pi}}_c)$$

lub dla dwóch klas k, c, gdy Z_k=Z_c:

$$b_{0k} + b_{1k}X_1 + b_{2k}X_2 + \cdots + b_{pk}X_p = b_{0c} + b_{1c}X_1 + b_{2c}X_2 + \cdots + b_{pc}X_p.$$

Ocena istotności modelu

Wyznaczone funkcje dyskryminacyjne podlegają ocenie istotności statystycznej. W tworzeniu funkcji dyskryminacyjnych dąży się do tego, by stosunek zmienności międzygrupowej do zmienności wewnątrz grup był jak największy. Właśnie to mierzy statystyka Lambda-Wilksa, której wzór jest następujący [15]:

$$\lambda = \frac{\det \boldsymbol{g}}{\det(\boldsymbol{g} + \boldsymbol{h})},$$

gdzie g – macierz wewnątrzgrupowej sumy kwadratów,

h – macierz międzygrupowej sumy kwadratów.

Statystyka przyjmuje wartości z przedziału <0, 1>. Jej wartość może być bezpośrednio interpretowalna – im niższa wartość, tym większa moc dyskryminacyjna modelu. Statystyka bliska zeru świadczy o doskonałej mocy dyskryminacyjnej, bliska jeden świadczy o zerowej mocy. Odpowiadający jej test F, którym bada się istotność statystyczną, posiada hipotezę zerową mówiącą o równości średnich grupowych. Wartość p poniżej przyjętego poziomu istotności świadczy o istotnej statystycznie dyskryminacji.

Statystyka Lambda-Wilksa służy do oceniania mocy dyskryminacyjnej całego modelu (łącznie wszystkich zmiennych). Jej bliźniacza postać może też zostać użyta do oceniania wkładu w dyskryminację osobno poszczególnych zmiennych. Wylicza się wtedy tzw. cząstkowy współczynnik Wilksa [16]:

$$\lambda_k^{cz} = \frac{\lambda'}{\lambda^0}$$

gdzie λ' – wartość współczynnika Lambda-Wilksa dla modelu po wprowadzeniu danej zmiennej,

λ^0 – wartość współczynnika Lambda-Wilksa dla modelu bez danej zmiennej.

Interpretacja współczynnika jest podobna jak statystyki Lambda-Wilksa. Przyjmuje wartości w przedziale <0, 1>; im niższa wartość tym badana zmienna wnosi więcej mocy dyskryminacyjnej do modelu.

Sprawdzenie spełnienia założeń LDA

Liniowa analiza dyskryminacyjna opiera się na założeniach:

- równości macierzy kowariancji grupowych,
- braku istotnej korelacji (współliniowości) między zmiennymi objaśniającymi,
- rozkładu normalnego danych.

W celu zbadania ostatniego założenia, bada się normalność rozkładu poszczególnych zmiennych. Jeśli wśród nich jest jedna zmienna, która znacząco odbiega od rozkładu normalnego, to prawdopodobnie cały model nie spełni tego założenia. Skutkuje to mniejszym lub większym spadkiem efektywności klasyfikacji. Jednakże wielu autorów, w tym P.A. Lachenbruch w pracy [17] udowodnili, że niewielkie odchylenia od rozkładu normalnego nie wpływają na jakość modelu. Najmocniejszym testem badającym normalność

rozkładu (z największym prawdopodobieństwem odrzuca hipotezę zerową, gdy ta jest fałszywa) jest test Shapiro-Wilka o następującej statystyce testowej [22]:

$$W = \frac{a_i(n)(X_{n-i+1} - X_i)^2}{(X_j - \bar{X})^2},$$

gdzie $a_i(n)$ są stałymi stablicowanymi zależnymi zarówno od liczebności próbki n oraz od i, a tzw. quasi-rozstępy rzędu i:

$$X_{n-i+1} - X_i = \begin{cases} i = 1, \dots, n, \text{ gdy } n \text{ parzyste,} \\ i = 1, \dots, n - \frac{1}{2}, \text{ gdy } n \text{ nieparzyste.} \end{cases}$$

Hipoteza zerowa zakłada, że rozkład badanej cechy jest rozkładem normalnym.

Założenie o braku współliniowości zmiennych objaśniających jest intuicyjne (nie dodaje się zmiennej do modelu, która nie wnosi żadnych nowych informacji) i występuje również w innych metodach klasyfikacji np. regresji logitowej. Współliniowość jest cechą zbioru danych statystycznych, pojawiającą się, gdy zmienne objaśniające są ze sobą silnie skorelowane. Wystąpienie dokładnej współliniowości uniemożliwia skonstruowanie poprawnego modelu. Często pojawia się wtedy problem łącznego braku istotności zmiennych objaśniających, estymatory tracą swoją efektywność, występują znaczne zmiany wartości szacowanych parametrów przy nieznacznym skróceniu lub wydłużeniu wielkości próby [20].

W analizie dyskryminacyjnej do oceny współliniowości danej zmiennej z innymi wykorzystuje się współczynnik tolerancji T_k zdefiniowany następująco:

$$T_k = 1 - R_k^2,$$

gdzie R_k^2 – współczynnik korelacji wielorakiej między daną zmienną, a pozostałymi zmiennymi występującymi w modelu. Współczynnik tolerancji określa część wariancji zmiennej, która nie jest wyjaśniana przez zmienne występujące razem z nią w modelu. Wprowadzanie do modelu zmiennych o niskich tolerancjach powoduje, że model staje się niedokładny [21].

Niespełnienie warunku równości macierzy kowariancji grupowych często nie skutkuje otrzymaniem niepoprawnych wyników. Mimo wszystko, nierówne macierze kowariancji powodują, że funkcje dyskryminacyjne są obciążone większym błędem. Potencjalnym źródłem błędów jest procedura szacowania wspólnej macierzy kowariancji (macierz $\sum$).

Zawiera ona estymatory wspólnych (równych) kowariancji grupowych w populacji. Macierz $\sum$ można estymować także wtedy, gdy kowariancje grupowe nie są równe, ale wówczas nie spełnia ona swoich podstawowych funkcji, polegających na upraszczaniu matematycznych formuł obliczeniowych. W rezultacie, szacunkowe prawdopodobieństwa przynależności grupowej będą obciążone większym błędem [18]. Najczęściej stosowanym testem do oceny równości macierzy kowariancji jest test M-Boxa. Box zaprezentował następującą statystykę testową [19]:

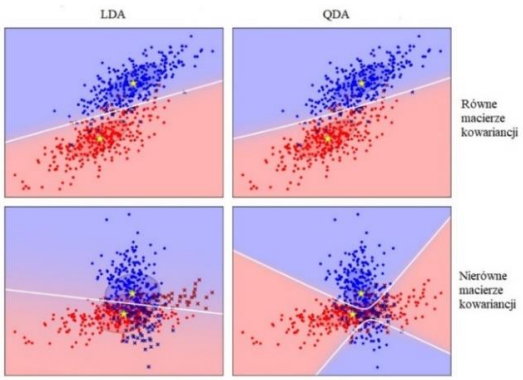
$$MB = (n - a)\ln|\sum_{s}| - \sum_{k=1}^{a}(n_k - 1)\ln|\sum_{k}|,$$

gdzie:

$\sum_{s}$ – międzygrupowa macierz kowariancji,

a – ilość klas (grup).

Statystykę MB można zatem traktować jako stosunek wyznacznika międzygrupowej macierzy kowariancji do średniej geometrycznej wyznaczników poszczególnych macierzy kowariancji grup. Wartość tego testu jest przekształcana na statystykę chi-kwadrat lub F z hipotezą zerową mówiącą o równości macierzy kowariancji. P-value poniżej przyjętego poziomu istotności świadczy o nierównych macierzach kowariancji. Do wyników testu należy podchodzić z pewną rezerwą, gdyż ma tendencje do odrzucania hipotezy zerowej w przypadku dużej próby oraz niewielkich nawet różnic w macierzach kowariancji grup. Jednakże duże różnice w macierzach kowariancji sprawiają, że LDA nie jest poprawną metodą dyskryminacji. W takiej sytuacji najczęściej stosuje się kwadratową analizę dyskryminacyjną (QDA). Test M-Boxa często jest stosowany na początku badań w celu ustalenia sposobu dalszej analizy – liniowej lub kwadratowej. Wpływ nierównych macierzy kowariancji na zachowanie się danych i wybór odpowiedniej metody dyskryminacji jest zilustrowany na rys. 11.

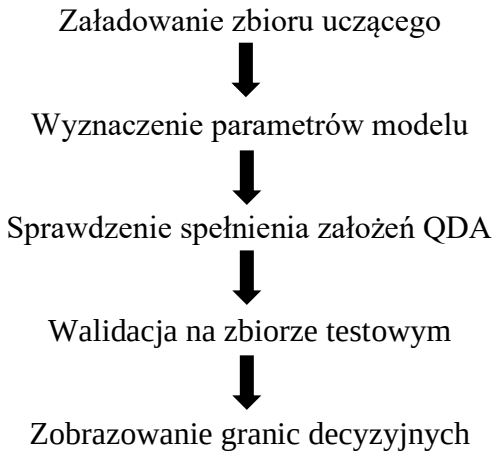


Rysunek 11. Macierze kowariancji grup, a metody LDA i QDA [43].

Walidacja na zbiorze testowym

W celu sprawdzenia poprawności klasyfikacji nowych danych posłużono się zbiorem testowym obserwacji, rozłącznym ze zbiorem uczącym. Na podstawie wyznaczonego modelu, nowe obserwacje zostały sklasyfikowane do poszczególnych klas. Porównano wyznaczone klasy z prawdziwymi etykietami klas i zaprezentowano wyniki w postaci tabeli klasyfikacyjnej. Takie same działania podjęto dla metod QDA, SVM, KNN oraz naiwnego klasyfikatora Bayesa.

2.2. Kwadratowa analiza dyskryminacyjna (QDA)



Wyznaczenie parametrów modelu

Podstawy teoretyczne QDA (ang. *Quadratic Discriminant Analysis*) są takie same jak LDA, z tym że w kwadratowej analizie dyskryminacyjnej zamiast wspólnej macierzy kowariancji $\sum$, przyjmuje się, że każda klasa posiada własną macierz kowariancji $\sum_k$. Obserwację przypisuje się do klasy, dla której funkcja:

$$\delta_{kq}(x) = -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{x}^T{\sum_k}^{-1}\boldsymbol{x}+\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{x}^T{\sum_k}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}{\boldsymbol{\mu}_k}^T{\sum_k}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}\log|\sum_k|+\log(\hat{\boldsymbol{\pi}}_k)$$

przyjmuje wartość maksymalną [44].

Obliczone wartości dyskryminacyjne funkcji nie są już liniowe względem x (jak w przypadku LDA), tylko kwadratowe, stąd nazwa tej metody. Powierzchnia rozdzielająca grupy jest drugiego stopnia i na płaszczyźnie przybiera postać okręgu, elipsy, paraboli lub hiperboli. Kwadratowa analiza dyskryminacyjna jest bardziej elastyczna w użyciu, kosztem mniej

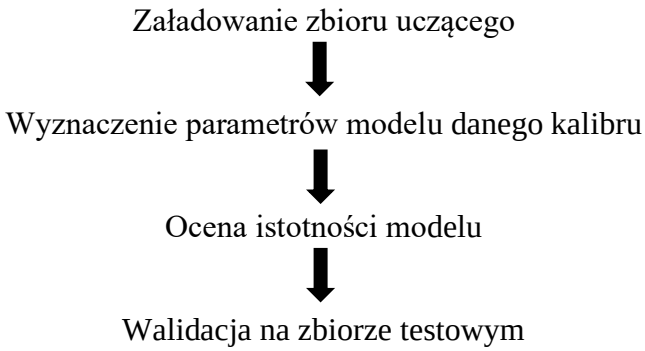
wygodnej interpretacji jej wyników. Nie jest możliwa interpretacja z najlepiej rozdzielającymi grupy kombinacjami liniowymi.

Sprawdzenie spełnienia założeń QDA

Klasyfikator QDA, podobnie jak LDA, wynika z założenia, że obserwacje z każdej klasy pochodzą z rozkładu normalnego. Rozkład sprawdza się w taki sam sposób. Nie jest wymagane założenie równości macierzy kowariancji w grupach.

2.3. Regresja logistyczna

Poniższy schemat postępowania powtarzany był osobno dla każdego z rozpatrywanych kalibrów.



Wyznaczenie parametrów modelu

Regresja logistyczna wywodzi się od regresji liniowej. W tym przypadku jednak modeluje się zmienną zero-jedynkową (dychotomiczną, dwumianową). Klasyczna regresja liniowa nie jest najlepszą metodą objaśniania zmiennej dychotomicznej, ponieważ wartości zmiennej objaśnianej obliczone przez model wykraczają poza zakres (0, 1), co powoduje często bardzo duże błędy predykcji.

Inne podejście polega na przyjęciu założenia, że rozważany jest model regresji:

$$Y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + u_i \ ,$$

w którym zmienna Y_i^* jest nieobserwowalna. Nazywa się ją zmienną ukrytą. Tym co obserwuje się jest zmienna zero-jedynkowa Y_i zdefiniowana jako:

$$Y_i = \begin{cases} 1, & \text{gdy } Y_i^* > 0 \\ 0, & \text{w pozostałych przypadkach.} \end{cases}$$

W tym przypadku zakłada się pewną zmienną ukrytą, dla której obserwuje się dychotomiczną realizację. Jeśli na przykład obserwowana zmienna zero-jedynkowa reprezentuje fakt, czy dana osoba jest zatrudniona, czy nie, wielkość Y_i^* byłaby zdefiniowana jako skłonność do znalezienia pracy [23].

Model logitowy jest szczególnym przypadkiem modelu zmiennej binarnej, w którym funkcja F opisująca prawdopodobieństwo ma postać:

$$p_i = F(x_i'\beta) = \Lambda(x_i'\beta) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

gdzie Λ to dystrybuanta rozkładu logistycznego.

Funkcja F jako dystrybuanta rozkładu logistycznego ma kształt krzywej typu S, co eliminuje problem pojawiający się w przypadku funkcji liniowej, że wszystkie wartości p_i, także te oszacowane, znajdują się w przedziale (0, 1). O praktycznym powodzeniu modelu logitowego zdecydował fakt, że funkcję odwrotną do F można zapisać jako:

$$x_i'\beta = Z_i = F^{-1}(p_i) = ln\frac{p_i}{1 - p_i}$$

Zatem zamiast modelować p_i względem zmiennych X, wygodnie jest modelować wyrażenie $ln\frac{p_i}{1-p_i}$ jako liniową funkcję tych zmiennych. To wyrażenie nazywa się logitem, a cały model nazywa się modelem logitowym (logistycznym). Logit to logarytm ilorazu szans, który jest definiowany jako $\frac{p_i}{1-p_i}$, czyli stosunek szans wystąpienia danego zdarzenia do szansy jego nie wystąpienia. Jeśli szanse są jednakowe (p_i = 0,5), to logit równa się zeru. Dla p_i < 0,5 logit jest ujemny, a dla p_i > 0,5 jest dodatni. Logitowa transformacja prawdopodobieństwa pozwala zastąpić wartość p_i przez liczbę z przedziału (-∞, +∞). Jest to główna korzyść płynąca z posługiwania się modelem logitowym [24].

Parametry w modelu logitowym szacowano metodą największej wiarygodności. Podstawowym pojęciem występującym w tej metodzie jest pojęcie wiarygodności próby i funkcji wiarygodności. Wiarygodnością n - elementowej próby prostej dla populacji o rozkładzie ciągłym przyjęto wartość wyrażenia:

$$L(x, \theta) = \prod_{t=1}^{n} f(x_t, \theta)$$

gdzie:

f – oznacza funkcję gęstości rozkładu prawdopodobieństwa,

$x = [x_1, x_2 \dots x_n]$ – próbę statystyczną,

θ – parametr bądź wektor parametrów.

Dla ustalonego wektora x wyników próby wiarygodność jest funkcją wartości parametru θ, nazywaną funkcją wiarygodności. Metoda największej wiarygodności polega na poszukiwaniu takich wartości θ, przy których funkcja L osiąga maksimum [25].
Niech zapisany będzie model logitowy w postaci:

$$ln\frac{p_i}{1-p_i} = Z_i = x_i'\beta = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Przedmiotem estymacji w tym modelu są parametry $\beta_0, \beta_1, \beta_2 \dots \beta_k$ będące elementami wektora β. Szacuje się je metodą największej wiarygodności. Dla niezależnej próby $Y_1, Y_2 \dots Y_n$ (gdzie Y_i = 1 lub 0 dla i = 1,2 … n), a prawdopodobieństwo P(Y_i = 1) = p_i, prawdopodobieństwo obserwowania jakiejś wartości Y_i, czy jest ona równa 1 czy 0, można zapisać łącznie jako P(Y_i) = ${p_i}^{Y_i}$ $(1-p_i)^{1-Y_i}$. Prawdopodobieństwo obserwowania konkretnej próby jest iloczynem prawdopodobieństw P(Y_i), czyli $\prod_{i=1}^{n} {p_i}^{Y_i}$ $(1-p_i)^{1-Y_i}$. Jest to funkcja wiarygodności.

W celu znalezienia jej maksimum wygodniej jest zmaksymalizować logarytm funkcji wiarygodności:

$$lnL = \sum_{i=1}^{n} [\, Y_i \, ln(p_i) + (1 - Y_i) \ln (1 - \, p_i)]$$

Dla modelu logitowego logarytm funkcji wiarygodności ma postać:

$$lnL = \sum_{i=1}^{n} [Y_i Z_i - \ln(1 + e^{Z_i})]$$

W wyniku maksymalizacji lnL otrzymuje się oceny $b_0, b_1, b_2 \dots b_k$. Funkcję wiarygodności jako funkcję tych ocen można zapisać następująco:

$$ln\hat{L} = \sum_{i=1}^{n} [Y_i \hat{Z}_i - \ln\left(1 + e^{\hat{Z}_i}\right)] \,,$$

gdzie $\hat{Z}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$. Maksimum funkcji jest osiągane w punkcie, w którym pierwsze pochodne lnL względem parametrów $b_0, b_1, b_2 \dots b_k$ równają się zeru.

Te równania noszą nazwę równań wiarygodności i są następujące:

$$\frac{\partial ln\hat{L}}{\partial b_j} = \sum_{i=1}^{n} [\, Y_i - \frac{e^{\hat{Z}_i}}{1 + e^{\hat{Z}_i}} \,] X_{ji} = 0 \qquad \text{dla } j = 0,1, \dots, k \text{ (przy tym } X_{0i} = 1)$$

Problem numeryczny polega na znalezieniu $k + 1$ rozwiązań układu równań wiarygodności. Równania te są nieliniowe względem poszukiwanych ocen parametrów $b_0, b_1, b_2 \dots b_k$. Powyższy układ równań ma postać:

$$\sum_{i=1}^{n} (Y_i - \hat{p}_i) X_{ji} = 0 \qquad \text{dla } j = 0,1, \dots, k$$

Oznacza to, że dla ocen parametrów modelu logitowego ważona suma odchyleń wartości Y_i od oszacowanych wartości prawdopodobieństwa p_i równa się zeru. Wagami są obserwowane w próbie wartości zmiennych X. Co więcej, jeśli tylko w modelu występuje wyraz wolny ($X_{0i} = 1$), to z powyższego równania wynika, że dla $j = 0$ średnia wartość Y jest równa średniej wartości oszacowanego prawdopodobieństwa p. Średnia wartość Y to proporcja wartości Y = 1 w próbie [24].

Wynikiem estymacji przeprowadzonej metodą największej wiarygodności jest wartość funkcji wiarygodności, macierz wariancji-kowariancji dla oszacowanych współczynników regresji (przydatna przede wszystkim przy obliczaniu przedziałów ufności) oraz lista zmiennych w modelu z odpowiadającymi oszacowanymi współczynnikami regresji oraz wartościami błędów standardowych.

Ocena istotności modelu

Do oceny istotności łącznie wszystkich zmiennych, a więc jakości całego modelu zastosować można m. in. testy ilorazu wiarygodności, dopasowania chi-kwadrat lub Hosmera-Lemeshowa. Poniżej opisano test ilorazu wiarygodności (MNW), który jest najczęściej wykorzystywany w pakietach statystycznych.

Estymatory MNW mają asymptotyczny rozkład normalny. W związku z tym test istotności dla pojedynczego parametru opiera się na statystyce o rozkładzie N(0, 1). Istotność całego modelu weryfikuje się za pomocą testu ilorazu wiarygodności. Hipoteza zerowa mówi, że wszystkie parametry przy zmiennych równają się 0. Statystyka testu ma postać:

$$LR = 2(lnL_{UR} - lnL_R),$$

gdzie:

L_{UR} – wartość funkcji wiarygodności dla pełnego modelu,

L_R – wartość funkcji wiarygodności dla modelu zawierającego tylko wyraz wolny.

Statystyka LR ma rozkład chi-kwadrat z liczbą stopni swobody równą liczbie wszystkich zmiennych objaśniających [26].

Jeśli ogólnie model jest skonstruowany poprawnie, następnym krokiem jest analiza istotności poszczególnych zmiennych i usunięcie tych nieistotnych. Istnieje kilka różnych testów zaprojektowanych do oceny znaczenia zmiennej niezależnej, opisany niżej jest test statystyki Walda, który jest podstawą oceny parametrów w większości pakietów statystycznych.

Statystyka Walda może być wykorzystywana do oceny istotności poszczególnych współczynników regresji. Jest to stosunek kwadratu współczynnika regresji do kwadratu błędu standardowego współczynnika:

$$\frac{\beta_i^2}{(SE\beta_i)^2}$$

Statystyka Walda ma asymptotyczny rozkład chi-kwadrat z jednym stopniem swobody. Zaletą tej metody jest jej łatwa kalkulacja [26].

Walidacja modelu

Jakość modelu oceniono na postawie tablicy trafności (tabeli klasyfikacyjnej), z uwagi na podobieństwo do tabeli klasyfikacyjnej wykorzystanej w innych metodach zastosowanych w pracy. Tablica trafności jest wynikiem krzyżowej klasyfikacji zmiennej wynikowej Y ze zmienną dychotomiczną, której wartości pochodzą z oszacowanych prawdopodobieństw logistycznych (tzw. pochodną zmienną dychotomiczną). Aby uzyskać pochodną zmienną dychotomiczną trzeba zdefiniować punkt odcięcia *c* i porównać każde oszacowane prawdopodobieństwo z *c*. Jeśli oszacowane prawdopodobieństwo przekracza *c*, to pochodna zmienna dychotomiczna będzie równa 1; w przeciwnym razie będzie równa 0. Najczęściej stosowana wartość dla *c* wynosi 0,5 [27]. Taką wartość zastosowano również w pracy. Postać tabeli klasyfikacyjnej dla regresji logistycznej przedstawiono jako tabelę 2.

Tabela 2. Postać tabeli klasyfikacyjnej dla metody regresji logistycznej.

Teoretyczne wartości	Empiryczne wartości	
	0	1
0	C	D
1	B	A

A, B, C, D to liczby całkowite oznaczające ilości obserwacji w każdej grupie:

A – oznacza liczbę zaobserwowanych jedynek (czyli wartości zmiennej dychotomicznej = 1), które zostały prognozowane przez model poprawnie jako jedynki,

B – oznacza liczbę zaobserwowanych jedynek, które zostały niepoprawnie prognozowane przez model jako zera,

C – oznacza liczbę zaobserwowanych zer, które zostały prognozowane przez model niepoprawnie jako jedynki,

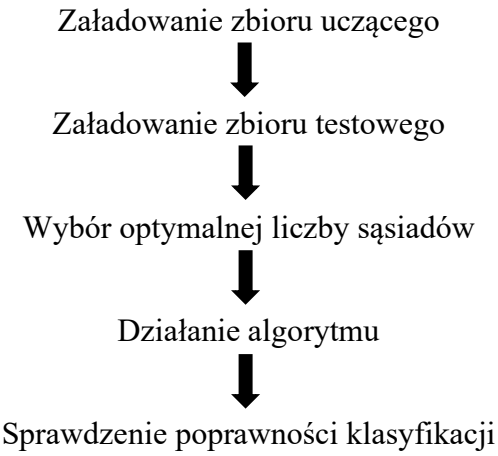
D – oznacza liczbę zaobserwowanych zer, które zostały prognozowane przez model poprawnie jako zera.

Najważniejszą miarą związaną z tabelą klasyfikacji jest tzw. zliczeniowy R^2. Jest to stosunek sumy poprawnie prognozowanych zer i jedynek do wszystkich obserwacji w modelu. Miara ta jest często stosowana do porównywania jakości modeli o różnych specyfikacjach. Wartości powyżej 0,7 świadczą o akceptowalnym dopasowaniu modelu do danych [46].

Regresja logistyczna nie wymaga spełnienia rygorystycznych warunków odnośnie poziomu dopasowania do danych, by oszacowania były wiarygodne. Występują tu typowe wymagania, występujące także w innych modelach, tj.: brak współliniowości predyktorów oraz niezależność obserwacji. Współliniowość zmiennych sprawdzono podczas analizy dyskryminacyjnej.

Na dokładność oszacowania modelu logitowego duży wpływ ma liczba obserwacji; uczenie maszynowe wymaga przynajmniej 10 przypadków na każdą zmienną niezależną, statystycy rekomendują przynajmniej 30 obserwacji na każdy szacowany parametr [47]. Zbyt mała liczba obserwacji oraz dobrze odseparowane klasy mogą spowodować, że parametry modelu logistycznego okażą się niestabilne [48].

2.4. Metoda K – najbliższych sąsiadów (KNN)



Działanie algorytmu

KNN (ang. *K-nearest neighbours*) jest przykładem tzw. leniwego klasyfikatora (ang. *lazy learner*). Nazwa leniwy nie odnosi się do prostoty algorytmu, lecz do tego, że nie uczy się on funkcji dyskryminacyjnej na podstawie danych uczących, lecz stara się „zapamiętać" cały zbiór próbek.

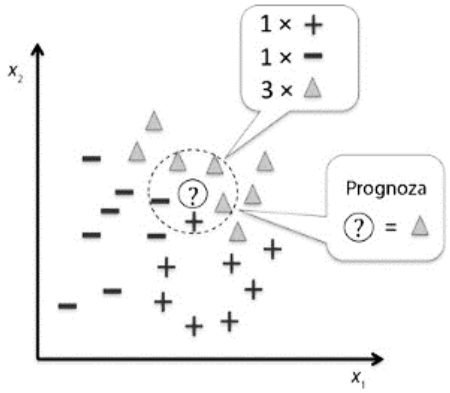
Działanie algorytmu KNN można opisać w następujących krokach [13]:

(1) Wybierz jakąś wartość parametru *K* i metrykę odległości.

(2) Znajdź *K* najbliższych sąsiadów próbki, którą chcesz sklasyfikować.

(3) Przydziel etykietę klasy poprzez głosowanie większościowe.

Zobrazowanie działania algorytmu przedstawiono na rys. 12:



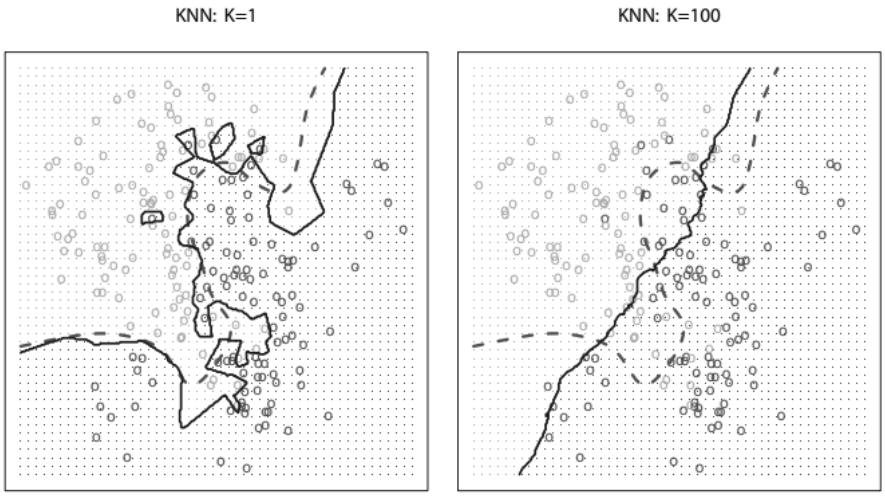
Rysunek 12. Działanie algorytmu KNN [13].

Na podstawie wybranej metryki odległości algorytm KNN wyszukuje w zestawie danych uczących *K* próbek znajdujących się najbliżej klasyfikowanego punktu lub wykazujących największe podobieństwo do niego. Etykieta klasy tej próbki zostaje określona poprzez większościowe głosowanie przeprowadzone pomiędzy K najbliższych sąsiadów. Największą zaletą takiego pamięciowego algorytmu jest natychmiastowe adaptowanie się klasyfikatora w trakcie pobierania nowych danych uczących. Równoważy to jednak główna wada, polegająca na liniowym wzroście złożoności obliczeniowej wraz z liczbą próbek uczących. Poza tym nie można odrzucać żadnych danych uczących, ponieważ w tym algorytmie nie występuje proces uczenia. Zatem w przypadku dużych zbiorów danych pojawia się problem z pojemnością nośników. Należy wspomnieć, że algorytm KNN jest bardzo wrażliwy na przetrenowanie z powodu tzw. „klątwy wymiarowości". Jest to zjawisko, w którym przestrzeń cech wraz ze wzrostem liczby wymiarów zestawu danych uczących o ustalonym rozmiarze staje się coraz bardziej rozległa. W wielowymiarowej przestrzeni nawet najbliżsi sąsiedzi znajdują się zbyt daleko, aby uzyskać za ich pomocą dobre oszacowanie [13].

Wybór liczby K sąsiadów

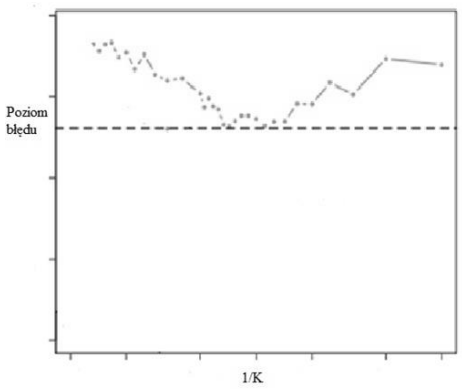
Właściwy dobór parametru *K* stanowi podstawę w uzyskaniu równowagi pomiędzy przetrenowaniem i zbyt małym dopasowaniem. Trzeba się także upewnić, że wybiera się metrykę odległości dopasowaną do cech zestawu danych. Często dla próbek przyjmujących wartości liczb rzeczywistych stosowana jest prosta metryka euklidesowa [13].

Dla *K*=1, krzywa rozgraniczająca dwie klasy najczęściej jest bardzo elastyczna. Wraz ze zwiększaniem K, granica przyjmuje coraz bardziej liniową postać (rys. 13):



Rysunek 13. Postać granicy między klasami a liczba *K* [28].

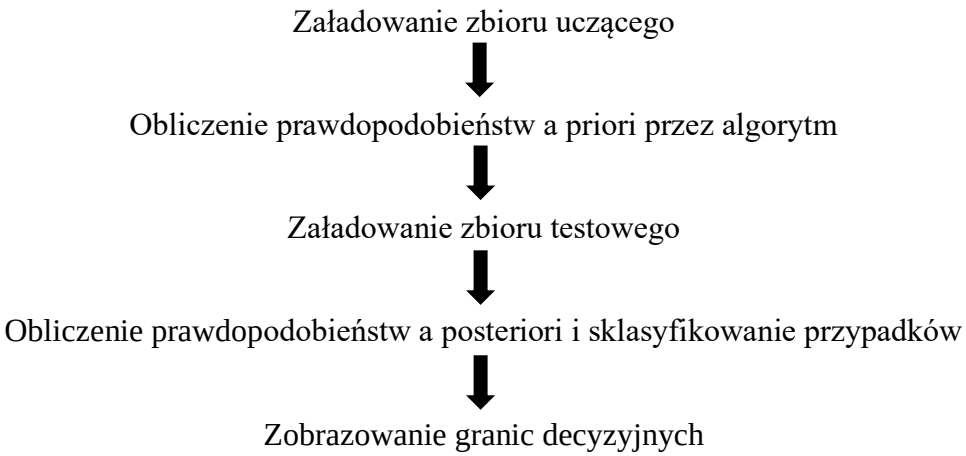
Otrzymanie liniowej postaci granicy między klasami nie oznacza, że uzyskane rozróżnienie jest najlepsze. Na rys. 14 przedstawiony jest wykres poziomu błędu klasyfikacji dla typowych danych.



Rysunek 14. Liczba K a poziom błędu klasyfikacji dla przykładowych danych [28].

Z rys. 14 wynika, że poziom błędu maleje tylko do pewnego momentu zwiększania liczby *K*, dlatego wybór właściwej liczby sąsiadów jest kluczowy do otrzymania jak najlepszych wyników.

2.5. Naiwny klasyfikator Bayesa



Naiwny klasyfikator Bayesa jest metodą klasyfikacji powstałą na bazie teorii Bayesa. Określenie „naiwny” odnosi się do faktu, iż w zastosowanym modelu prawdopodobieństwa przyjęto pełną niezależność zmiennych losowych, a to w rzeczywistości często mija się z prawdą. Pomimo to, naiwny klasyfikator Bayesa daje zaskakująco dobre rezultaty.

W myśl teorii Bayesa, prawdopodobieństwo, że zaobserwowany przypadek o cechach $X_1, \dots, X_n$, należy do klasy C, wyraża się wzorem [29]:

$$P(C|X_1, \dots, X_n) = \frac{P(C)P(X_1, \dots, X_n|C)}{P(X_1, \dots, X_n)}$$

Mianownik w powyższym równaniu nie zależy od C, a ponadto dane są wartości cech X_i, zatem przyjmuje się, że mianownik ten jest wartością stałą. Licznik można przedstawić jako:

$$P(C, X_1, \dots, X_n) = P(C)P(X_1, \dots, X_n|C) = P(C)P(X_1|C)P(X_2, \dots, X_n|C, X_1)$$
$$= P(C)P(X_1|C)P(X_2|C, X_1)P(X_3, \dots, X_n|C, X_1, X_2) = \cdots$$

Z „naiwnego" założenia, że nie istnieje warunkowa zależność pomiędzy X_i oraz X_j:

$$P(X_i|C, X_j) = P(X_1|C)$$

Zatem

$$P(C, X_1, \dots, X_n) = P(C)P(X_1|C)P(X_2|C)P(X_3|C) \dots$$

Stąd

$$P(C|X_1, \dots, X_n) = Z * P(C) \prod_{i=1}^{n} P(X_n|C),$$

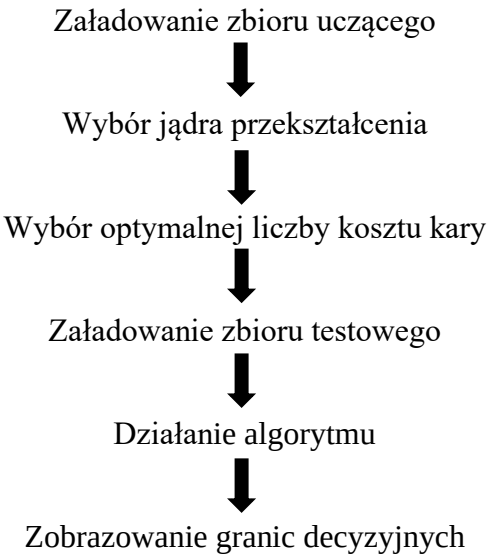
gdzie Z – współczynnik skali zależny tylko od $X_1, \dots, X_n$ (stała, jeżeli wartości cech są znane) [29].

Naiwny klasyfikator Bayesa łączy przedstawiony model prawdopodobieństwa z regułą decyzyjną. Regułą tą jest MAP (ang. *maximum a posteriori*) [29]. Przypadek jest więc sklasyfikowany do klasy z największym prawdopodobieństwem *a posteriori* (obliczanym według wzorów przedstawionych powyżej jako iloczyn prawdopodobieństwa *a priori* i szansy, że obserwacja będzie należeć do danej grupy). W celu obliczenia prawdopodobieństwa *a priori* wykorzystuje się częstości występowania przypadków.

Obliczenie szansy następuje na podstawie rozkładu normalnego, lognormalnego, gamma lub Poissona w zależności od konkretnej sytuacji [29].

Rezultaty działania naiwnego klasyfikatora Bayesa wydają się zatem być podobne do tych otrzymywanych przez LDA. Tak jednak nie jest, ponieważ założenie o niezależności predyktorów w każdej klasie powoduje, że macierz kowariancji $\sum_k$ przyjmuje postać diagonalną [44].

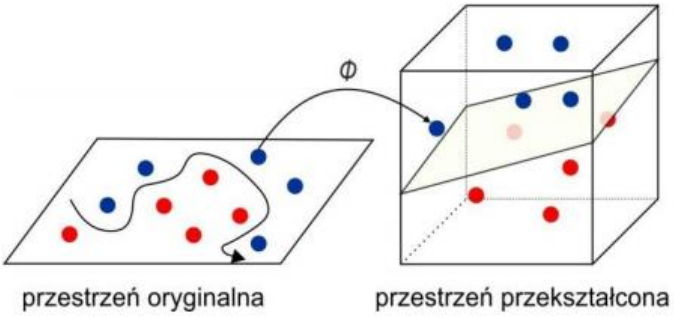
2.6. Metoda wektorów nośnych (SVM)



Działanie algorytmu

Metoda wektorów nośnych (*ang. Support Vector Machine*) jest metodą uczenia maszynowego, wykorzystywaną głównie do klasyfikacji danych. Jest to jedyna metoda w uczeniu maszynowym wypracowana „od teorii do praktyki". Wszystkie inne algorytmy były najpierw odkrywane dzięki eksperymentom, dopiero później została opisywana teoria działania. W algorytmie SVM każdy element danych jest wykreślany jako punkt w przestrzeni, zależnej od zastosowanej funkcji jądrowej. Następnie dokonuje się klasyfikacji poprzez znalezienie optymalnej hiperpłaszczyzny oddzielającej dane [41].

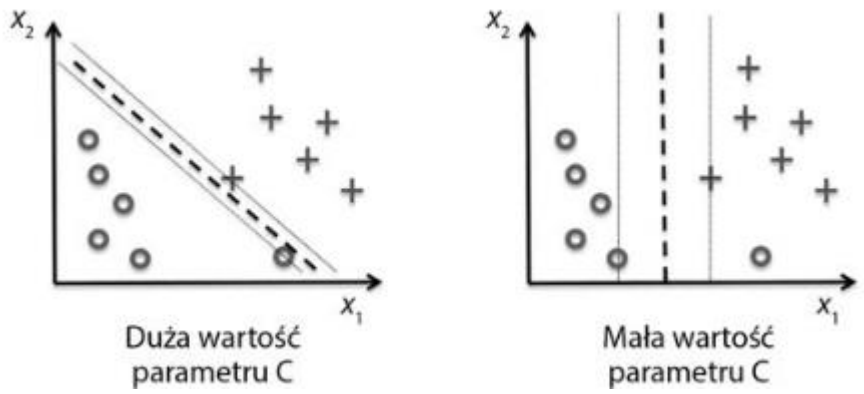
Hiperpłaszczyzna jest budowana w taki sposób, by oddzielać klasy z maksymalnym możliwym marginesem. Dane znajdujące się najbliżej hiperpłaszczyzny nazywane są wektorami nośnymi (ang. *support vectors*) i są najbardziej istotnymi elementami klasyfikacji. Wynik jest zależny od szerokości marginesu; im margines jest szerszy, tym niższy jest błąd klasyfikacji. W metodzie SVM do przekształcania przestrzeni wykorzystywane są funkcje jądrowe liniowe i radialne. Rzadziej stosowane są funkcje jądrowe wielomianowe i sigmoidalne. Na rys. 15 zobrazowane jest przekształcanie punktów w przestrzeni [30].



Rysunek 15. Przekształcenie przestrzeni w metodzie SVM [30].

Wybór parametru kosztu kary

W przypadku każdej z funkcji jądrowych istnieje możliwość zdefiniowania tzw. parametru C, nazywanego kosztem kary (ang. *cost of penalty*), który odpowiada za kontrolę kompromisu pomiędzy błędami w klasyfikacji, a wymuszaniem marginesów. Wysokie wartości parametru świadczą o wysokiej karze nakładanej na obserwacje znajdujące się po niewłaściwej stronie hiperpłaszczyzny lub wewnątrz marginesu. Spada przez to skuteczność klasyfikacji. Niskie wartości parametru są przyczyną większej wariancji, czyniąc klasyfikator nieużytecznym. Nieprawidłowo sklasyfikowane obserwacje mogą być uznawane przez algorytm za poprawne [30]. Zależność między parametrem C, a wielkością marginesu jest przedstawiona na rys. 16:



Rysunek 16. Koszt kary, a szerokość marginesu [13].

Metoda wektorów nośnych przez wielu badaczy jest uważana za najdokładniejszą spośród wszystkich metod uczenia maszynowego oraz za szybszą w implementacji od algorytmów sieci neuronowych [30].

2.7. Porównanie metod klasyfikacji

Regresja logistyczna oraz liniowa analiza dyskryminacyjna są ze sobą powiązane. Rezultatem zastosowania obydwu metod jest wyznaczenie liniowych funkcji klasyfikacyjnych. Jedyna różnica między tymi podejściami jest taka, że współczynniki w regresji logistycznej są wyznaczone za pomocą metody największej wiarygodności, podczas gdy w LDA są obliczone na podstawie średnich i wariancji rozkładu normalnego. W związku z tym można się spodziewać, że obydwie metody dadzą podobne rezultaty. W praktyce nie zawsze ma to miejsce. LDA zakłada, że obserwacje przyjmują rozkład normalny ze wspólną macierzą kowariancji w każdej klasie, zatem może być bardziej wiarygodną i skuteczną metodą, kiedy te założenia są w przybliżeniu utrzymane. Jeśli założenia te nie są spełnione, wtedy skuteczniejszą metodą powinna być regresja logistyczna, która nie ma tak surowych warunków do spełnienia. Jednakże, jeżeli klasy są dobrze odseparowane od siebie, oszacowane parametry w regresji logistycznej mogą być niestabilne [28].

Metody uczenia maszynowego: K-najbliższych sąsiadów, wektorów nośnych oraz naiwny klasyfikator Bayesa prezentują inne, nieparametryczne podejście. W metodzie KNN nie przyjmuje się żadnych założeń dotyczących kształtu granicy decyzyjnej. Obserwacja jest przypisywana do klasy, w której już znajdują się najbardziej podobne do niej elementy. Naiwny klasyfikator Bayesa jest oparty na teorii Bayesa i przypisuje obserwację do klasy z największym prawdopodobieństwem *a posteriori*, będącego iloczynem *a priori* i indywidualnej szansy występowania obiektu w danej klasie. Z kolei metoda wektorów nośnych przekształca punkty do odpowiedniej przestrzeni i dopiero wtedy wyznacza separującą hiperpłaszczyznę. Wszystkie te metody prezentują zupełnie inne podejście do klasyfikacji, dlatego analiza ich wyników jest interesująca.

Ostatnią opisaną metodą klasyfikacji jest metoda QDA. Jest ona kompromisem między nieparametrycznym metodami uczenia maszynowego oraz parametrycznymi metodami LDA i regresji logitowej. Ponieważ QDA zakłada kwadratową funkcję rozróżniającą klasy, może dokładnie modelować szerszy zakres problemów niż metody liniowe. Pomimo, że nie jest tak elastyczna jak metody nieparametryczne, to może lepiej klasyfikować dane, ponieważ przyjmuje pewne założenia formy funkcji rozróżniających. Wadą tej metody w stosunku do regresji logitowej i LDA jest to, że nie generuje bezpośrednio interpretowalnych współczynników funkcji.

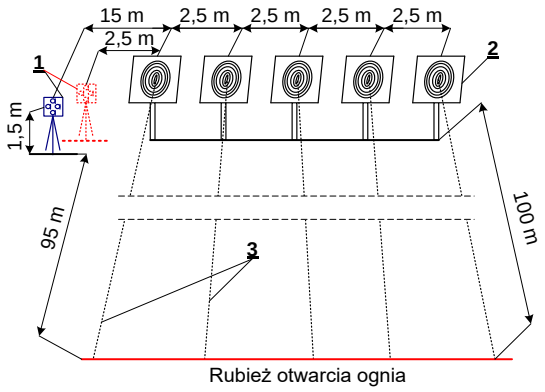
Z powyższych rozważań wynika, że nie ma zatem jednej najlepszej metody klasyfikacji. Jej wybór zależy od posiadanych danych. Jeśli spodziewać się można, że granice rozróżniające klasy są liniowe, wtedy najlepszym wyborem wydaje się być regresja logistyczna, lub jeśli dodatkowo spełnione są założenia, to liniowa analiza dyskryminacyjna. Jeśli granica między danymi jest delikatnie nieliniowa, zmierzając w kierunku funkcji kwadratowej, wtedy metoda QDA powinna być odpowiednia (przy założeniu, że dane nie odbiegają znacząco od rozkładu normalnego). W przypadku danych o granicach elastycznych, bez regularnego kształtu, metody uczenia maszynowego są najlepszym wyborem, gdyż zostały stworzone do tego, by efektywnie klasyfikować każdy zbiór elementów.

3. Proces zebrania i przygotowania danych

W tym rozdziale opisano proces pozyskania danych z badań oraz przedstawiono ich krótki opis.

Dane dla kalibru 7,62 mm

Jednym z badań, w którym gromadzono dane podczas strzelań z broni palnej kalibru 7,62 mm były badania opisane w opracowaniu [31]. W badaniach na strzelnicy WAT wykorzystano pociski 7,62x39 mm wz. 43. Przygotowane zostało stanowisko pomiarowe jak na rys. 17.



Rysunek 17. Stanowisko pomiarowe [31].

1-stanowisko umieszczenia mikrofonów, 2 – tarcze strzelnicze, 3-tor lotu pocisku.

Pomiary odbywały się w dwóch etapach: w pierwszym etapie cztery czujniki akustyczne (mikrofony) zostały rozstawione 2,5 m od środka pierwszej tarczy, a w drugim etapie badań odległość wynosiła 15 m. Czujniki umieszczone były w pomijalnie małej odległości w stosunku do siebie na wysokości trajektorii lotu pocisku. Strzelano z różnych stanowisk ogniowych. Liczba zarejestrowanych strzałów wyniosła 202. Badania prowadzono przy bezwietrznej pogodzie w temperaturze 20°C, przy ciśnieniu powietrza 1010 hPa i wilgotności 70 %, a więc w typowych warunkach atmosferycznych. Odległość lufy do tarczy wynosiła 100 m, co pozwoliło uniknąć zaburzenia rejestrowanych przebiegów fali N przez gazy wylotowe z lufy podczas strzału. W rezultacie strzelań zostały zgromadzone tylko przebiegi z zarejestrowanymi zaburzeniami ciśnienia pochodzącymi od pocisków. Nie były one przyporządkowane do poszczególnych strzałów, co było głównym problemem do rozwiązania podczas analizy tego zbioru danych.

Na podstawie otrzymanych przebiegów, wyznaczone zostały parametry fali N: wartość szczytowa amplitudy ciśnienia oraz czas trwania fali N. Do automatycznego wyznaczania parametrów fali N użyto skryptu napisanego w języku Python. Dla potwierdzenia poprawności wyznaczenia parametrów, ponownie w programie MS Excel zobrazowano

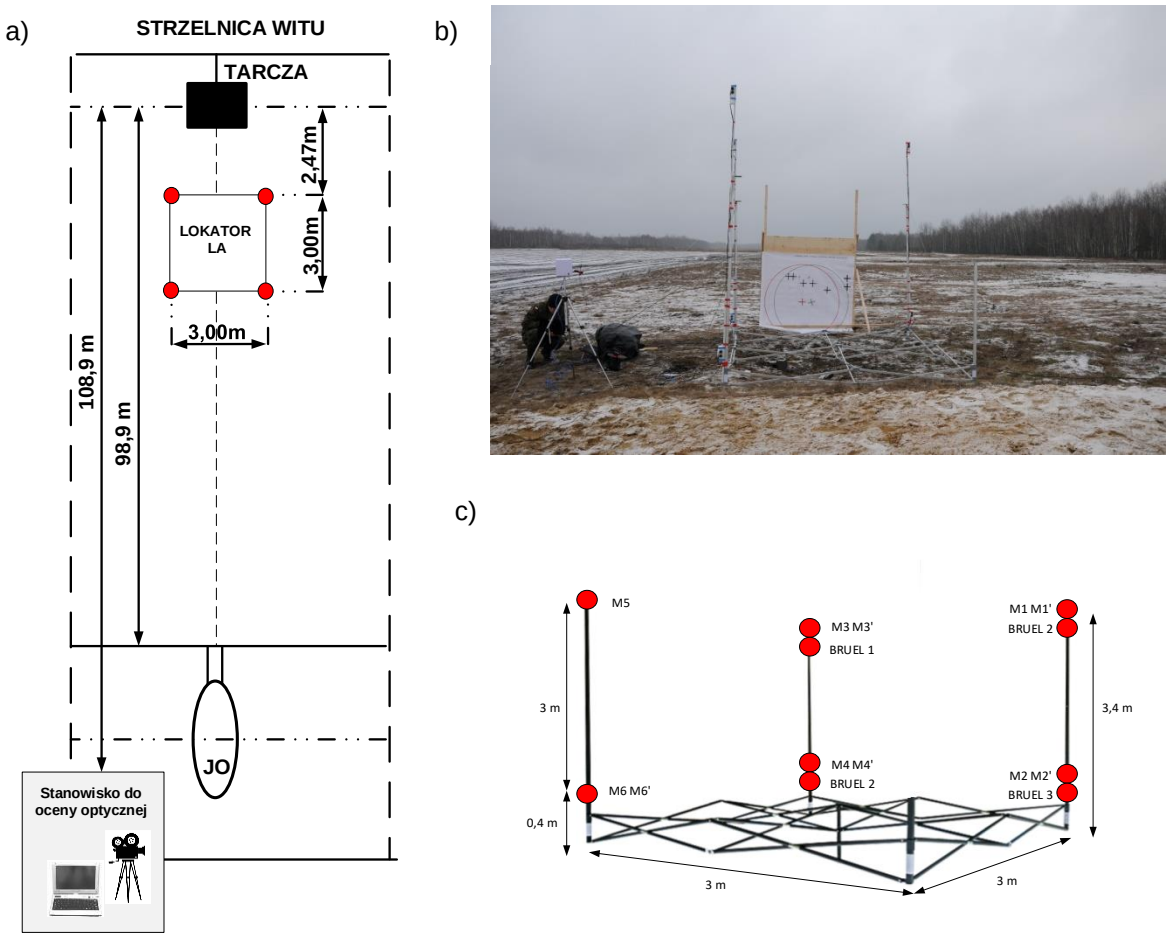
i wyznaczono parametry fali. Wartość szczytowa amplitudy ciśnienia została wykorzystana przy grupowaniu przebiegów do odpowiednich zakresów odległości czujników od trajektorii lotu pocisku (w przedziale od 2,5 do 25 [m] z krokiem co 2,5 m). Porównywano wartość amplitudy z wartością teoretyczną obliczoną przy pomocy modelu Whithama. W rezultacie analizy otrzymano 182 pliki z poprawnie zarejestrowaną falą N przyporządkowane do następujących odległości pomiarowych: 2,5; 7,5; 10; 15; 20; 25 [m]. Każdy strzał został zarejestrowany przez cztery czujniki z częstotliwością próbkowania 65 kHz, z uwagi jednak na to, że każdy czujnik rejestrował ten sam czas trwania, do modelu zostały wykorzystane dane z tylko jednego czujnika. Główną zaletą badań była duża liczba uzyskanych przebiegów dla różnych odległości rozstawienia mikrofonów pomiarowych od trajektorii lotu pocisku, co znacznie wzbogaciło zbiór danych.

Dane dla kalibrów 5,56 mm i 7,62 mm.

Podczas badań, opisanych w sprawozdaniu [32], przeprowadzonych przez pracowników Zespołu Radioelektroniki WAT w dniu 16.12.2006 r. wyznaczane zostały przebiegi fali N podczas strzelań z broni palnej kalibrów 5,56 mm oraz 7,62 mm. Badania te cechują się najwyższą wiarygodnością, zostały przeprowadzone w warunkach laboratoryjnych. Wykonanie badań w tunelu balistycznym Instytutu Techniki Uzbrojenia WML WAT umożliwiło redukcję hałasów zewnętrznych oraz skuteczną separację zaburzeń przebiegów fali N od gazów wylotowych i fali odbitej, które mogłyby wydłużać czas trwania fali N. Całkowicie wyeliminowany został wpływ wiatru i innych czynników atmosferycznych. Wszystkie pociski (7,62x39 mm wz. 43; 5,56x45 mm) były wystrzeliwane z nieruchomego karabinka ze stałą linią celowania skierowaną w środek tarczy. Mikrofony pomiarowe zostały ustawione na wysokości trajektorii lotu pocisku. Pomiar parametrów strzału zapewniały dwa mikrofony ustawiane w różnych odległościach: 0,35; 0,7; 0,85; 1,03; 1,22; 1,25 [m]. Oddano 23 strzały, ustawiając różne częstotliwości próbkowania. Dzięki dobrej dokumentacji badań, nie było problemów z przyporządkowaniem strzału do odległości ustawienia mikrofonu względem trajektorii pocisku. Analiza wyeksportowanych plików wykazała, że nie wszystkie rejestracje zawierają falę N, ponadto część z nich zawierała zbyt niską częstotliwość próbkowania (przyjęto, że dostateczny czas próbkowania nie może być mniejszy od 0,02 ms). Ostatecznie wyznaczono parametry 23 przebiegów fali N, w tym 13 dla kalibru 7,62 mm, a 10 dla kalibru 5,56 mm. Podsumowując niniejsze badania, mała liczba pomiarów jest rekompensowana przez bardzo dużą dokładność pomiarów.

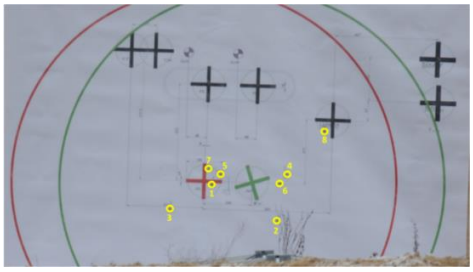
Dane dla kalibru 23mm

14 grudnia 2016 r. przeprowadzono badania [33], na terenie poligonu WITU w Zielonce, w celu weryfikacji poprawności przetwarzania zmian ciśnienia akustycznego wywołanego falą uderzeniową przez wybrane czujniki. Strzelano z armaty ZU23x2 pociskami 23x152 mm. Rezultatem tych badań były pliki z zarejestrowanymi przebiegami fali N. Stanowisko pomiarowe zostało ustawione zgodnie z rys. 18:



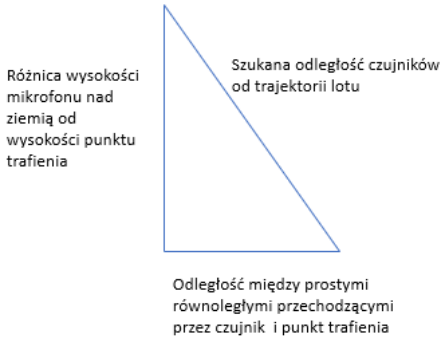
Rysunek 18. Stanowisko pomiarowe [33]: a) rozmieszczenie aparatury badawczej, b) widok apertury pomiarowej wraz z tarczą strzelniczą, c) rozmieszczenie czujników.

W sumie oddano 8 strzałów, które zostały zaznaczone na tarczy w sposób, jak na rys. 19:



Rysunek 19. Zobrazowanie przestrzelin po strzelaniach [33].

Zachowane zostały pliki zawierające rejestrację z mikrofonów oznaczonych jako M6, M5, M4 i M3, które były przedmiotem analiz. Wyznaczanie czasu trwania fali N odbywało się z pomocą programu PicoScope. Wyznaczone czasy trwania były bardzo dokładne, dzięki częstotliwości próbkowania równej 1 MHz. Nie zostały udokumentowane odległości mikrofonów od toru lotu pocisków. Zostały one wyznaczone dzięki znajomości punktów trafienia w cel, wymiarów tarczy oraz dokładnego rozstawienia stanowiska pomiarowego. Punkt celowania oraz rozstawienie czujników były stałe. Z powodu znacznej różnicy w wysokości rozlokowania poszczególnych mikrofonów oraz punktów trafienia w tarczę, szukaną odległość wyliczono z geometrii rozstawienia czujników pomiarowych względem trajektorii pocisków, w sposób zobrazowany na rys. 20.



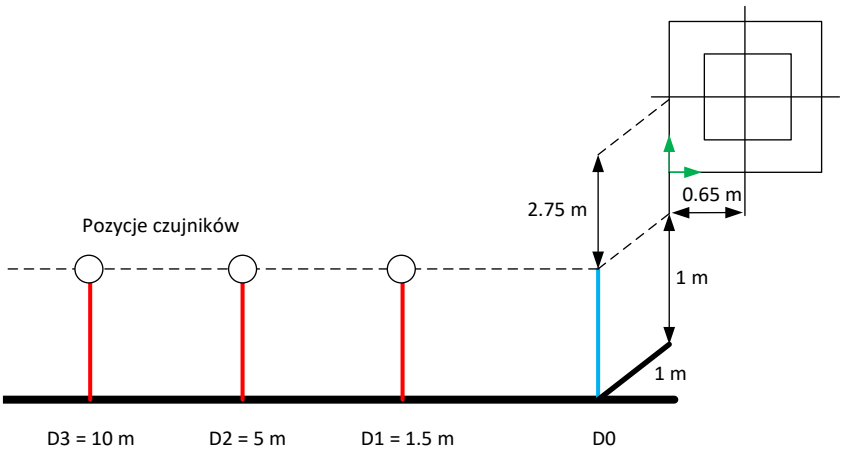
Rysunek 20. Geometria wyznaczenia odległości mikrofonów do toru lotu pocisku.

Mikrofony M6 i M4 oraz M3 i M5 zostały umieszczone parami na tej samej wysokości po tej samej stronie tarczy, w związku z czym ich odległość do trajektorii lotu była taka sama dla poszczególnych strzałów.

Rezultatem analiz było wyznaczenie czasu trwania fali N i odległości położenia czujników od toru lotu pocisku dla 32 zarejestrowanych strzałów.

Dane dla kalibru 35 mm

Dane dla kalibru 35 mm pochodzą z projektu [34] realizowanego na Wydziale Mechatroniki WAT. Przeprowadzono strzelania z armaty KDA na Centralnym Poligonie Sił Powietrznych w Ustce w dniach 9-10.01.2018. Wykorzystano naboje z pociskiem ćwiczebnym TP-T 35x228 mm. Niektóre serie strzelań (na odległość 100 m) zostały dobrze udokumentowane wraz z zaznaczeniem odległości stanowiska pomiarowego od toru lotu pocisku, w związku z czym zawierały użyteczne dane do modelu. Na rys. 21 i rys. 22 przedstawiono schemat wzajemnego rozmieszczenia mikrofonów pomiarowych oraz ich położenia względem tarczy.



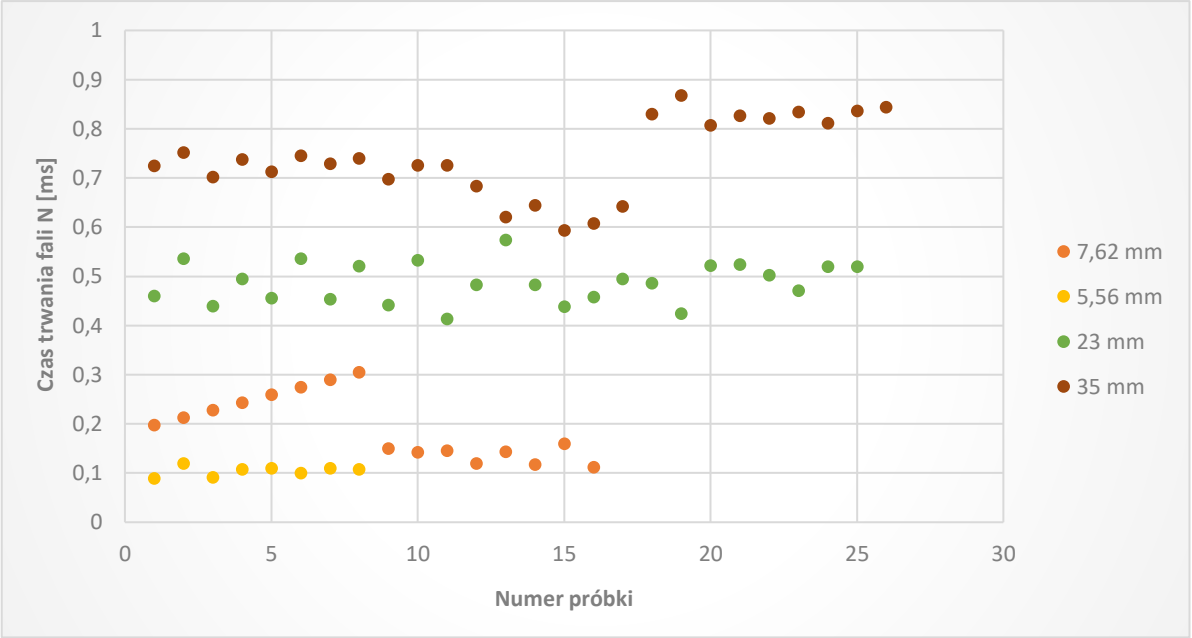
Rysunek 21. Schemat rozmieszczenia czujników względem tarczy [34].

Rysunek 22. Rozmieszczenie mikrofonów względem siebie [34].

W związku z tym, że każdy strzał posiadał informację o odległości, wyznaczenie zbioru danych nie było trudne. Strzały rejestrowane były z częstotliwością próbkowania równą 1 MHz. Czas trwania poszczególnych zaburzeń został odczytywany z wyeksportowanych podczas badań plików PicoScope. Otrzymano informacje o 42 przebiegach fali N.

Dobór zmiennych do modelu

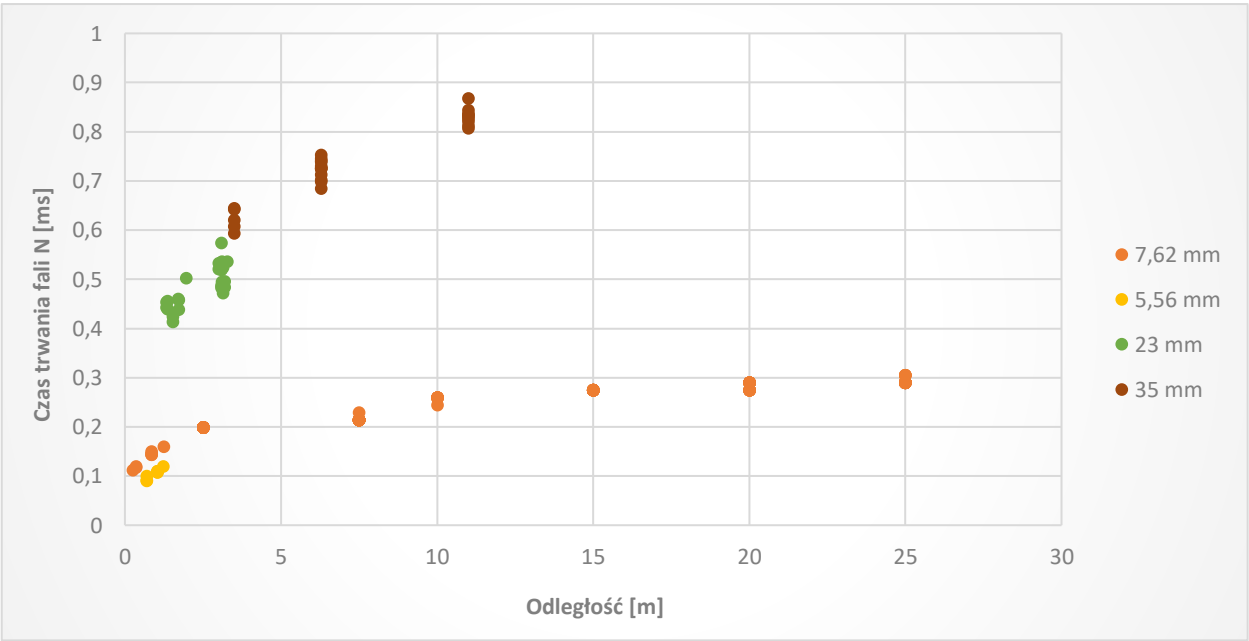
Na podstawie teoretycznych rozważań jako zmienną główną rozróżniającą pociski przyjęto czas trwania fali N dla każdego z rozpatrywanych kalibrów. Wykres danych empirycznych przedstawia poszczególne czasy trwania fali dla wszystkich kalibrów (rys. 23).



Rysunek 23. Czas trwania fali N.

Z wykresu przedstawionego na rys. 23 wynika, że czas trwania fali N jest dobrą zmienną grupującą pociski. Czasy trwania są dłuższe dla większych kalibrów. Grupy danych 35 i 23 [mm] są całkowicie rozłączne z grupami kalibrów 7,62 i 5,56 [mm]. Jednakże, w przypadku dwóch najmniejszych kalibrów wartości na siebie nachodzą, co wskazuje na konieczność dodania dodatkowych zmiennych do modelu.

Według modelu Whithama, czas trwania fali N zależy od odległość czujników pomiarowych od trajektorii pocisku. Sprawdzenie tej tezy dla danych empirycznych ilustruje wykres (rys. 24).



Rysunek 24. Rozkład czasu trwania fali N w funkcji odległości.

Rozkład czasu trwania fali N w funkcji odległości wykazuje dodatnią korelację między zmiennymi dla każdego rodzaju kalibru. Wskazuje to na możliwość dodania także tej zmiennej do modelu.

Ostatnią zmienną wyłonioną na postawie teoretycznych rozważań, jest prędkość pocisku w trakcie rejestracji. Z uwagi na brak danych empirycznych nie można było jej uwzględnić.

4. Budowa klasyfikatorów

Zastosowano w praktyce każdą z opisanych w rozdziale drugim metod. Dane podzielono losowo na dwa zbiory: uczący i testowy w proporcji 2:1, otrzymując 185 obserwacji dla zbioru uczącego oraz 93 dla zbioru testowego. Zestawienie danych zamieszczono w załączniku 2. Na podstawie zbioru uczącego nauczono każdy klasyfikator rozróżniać nowe obserwacje. Ich skuteczność sprawdzono na podstawie zbioru testowego. Modele zbudowano w całości za pomocą języka R, a prawie wszystkie statystyki, wykresy i wycinki przedstawione w niniejszym rozdziale pochodzą z tego środowiska. Do oceny istotności statystycznej modelu LDA wykorzystano pakiet Statistica. W przeprowadzonych analizach przyjęto następujące oznaczenia grup pocisków: **1** – oznacza amunicję kalibru 7,62 mm; **2** – kalibru 5,56 mm; **3** – kalibru 23 mm; **4** – kalibru 35 mm. Ponadto zmienna *kaliber* oznacza kaliber pocisku, zmienna czas oznacza czas trwania fali N w [ms], a zmienna *odl* oznacza odległość czujnika od trajektorii lotu pocisku w [m]. Kody źródłowe opracowanych algorytmów klasyfikacji, napisane w języku programowania R, zamieszczono w załączniku 1.

4.1. Model LDA

Oszacowane zostały następujące dyskryminacyjne liniowe funkcje Fischera:

$$Z_1 = -17{,}0737 + 174{,}3645 \cdot czas - 0{,}7329 \cdot odl$$

$$Z_2 = 8{,}1311 + 100{,}1279 \cdot czas - 0{,}5462 \cdot odl$$

$$Z_3 = -114{,}969 + 476{,}681 \cdot czas - 2{,}639 \cdot odl$$

$$Z_4 = -248{,}807 + 704{,}054 \cdot czas - 3{,}824 \cdot odl$$

W celu przyporządkowania obserwacji do odpowiedniej grupy, podstawiono wartości *odl* oraz *czas* do wszystkich czterech funkcji dyskryminacyjnych. W przypadku, gdy maksymalną wartość przyjmuje funkcja:

Z_1 – oznacza to, że pocisk jest kalibru 7,62 mm,

Z_2 – oznacza to, że pocisk jest kalibru 5,56 mm,

Z_3 – oznacza to, że pocisk jest kalibru 23 mm,

Z_4 – oznacza to, że pocisk jest kalibru 35 mm.

Zestaw wyznaczonych funkcji dyskryminacyjnych tworzy model identyfikujący pociski.

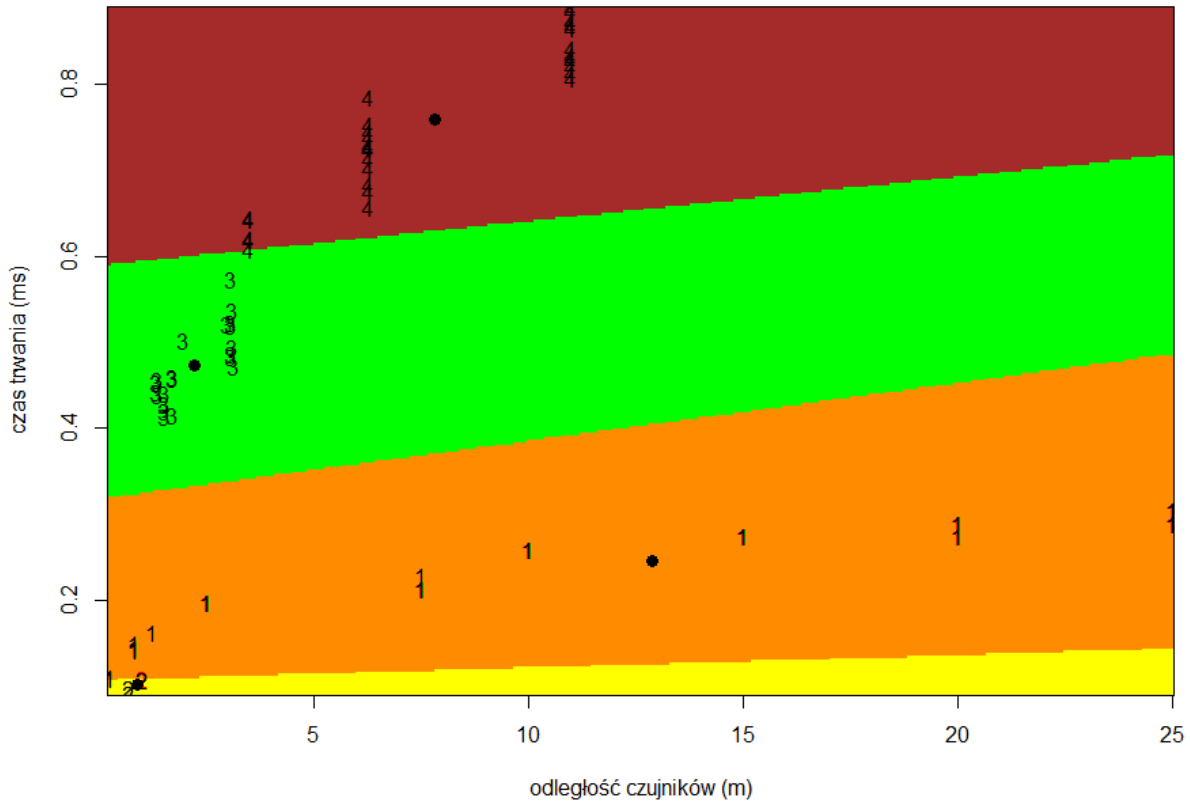
Walidacja modelu

Tabela klasyfikacyjna jest przedstawiona jako wyc. 1:

	empiryczne			
teoretyczne	1	2	3	4
1	60	0	0	0
2	2	4	0	0
3	0	0	12	0
4	0	0	0	15

Wycinek 1. Tabela klasyfikacyjna modelu LDA.

Model sklasyfikował poprawnie 96,78% przypadków. Dwie fale N pochodzące od pocisku 7,62 mm zostały sklasyfikowane jako pochodzące od kalibru 5,56 mm. Zobrazowanie sklasyfikowania danych wraz z liniowymi granicami jest przedstawione na rys. 25. Poprawnie sklasyfikowane przypadki są oznaczone na czarno, niepoprawnie na czerwono.



Rysunek 25. Klasyfikacja przypadków wraz z zaznaczonymi granicami.

Ocena istotności modelu

Wyniki testów istotności przeprowadzone w programie Statistica, przedstawione są na rys. 26. Przyjęto poziom istotności $\propto = 0{,}05$.

N=185 | Podsumowanie analizy funkcji dyskryminacyjnej. (Arkusz1)
Zmiennych w modelu: 2;Grupująca: kaliber (4 grup)
Lambda Wilksa: ,01981 przyb. F (6,360)=366,30 p<0,0000

	Lambda Wilksa	Cząstk. Wilksa	F usun. (3,180)	p	Toler.	1-Toler. (R-kwad)
czas	0,745125	0,026586	2196,846	0,00	0,382248	0,617752
odl	0,069273	0,285966	149,815	0,00	0,382248	0,617752

Rysunek 26. Podsumowanie analizy dyskryminacyjnej w programie Statistica.

Przedstawiony na rys. 26 fragment tabeli wyników z programu Statistica zawiera wszystkie niezbędne informacje o istotności zarówno całego modelu, jak i jego zmiennych.

Otrzymana wartość statystyki Lambda-Wilksa dla modelu równa 0,01981 wraz z wartością p dążącą do zera oznacza, że model w sposób istotny statystycznie rozróżnia grupy. Ponadto wartość statystyki bliska zeru świadczy o tym, że model posiada bardzo dobrą moc dyskryminacyjną.

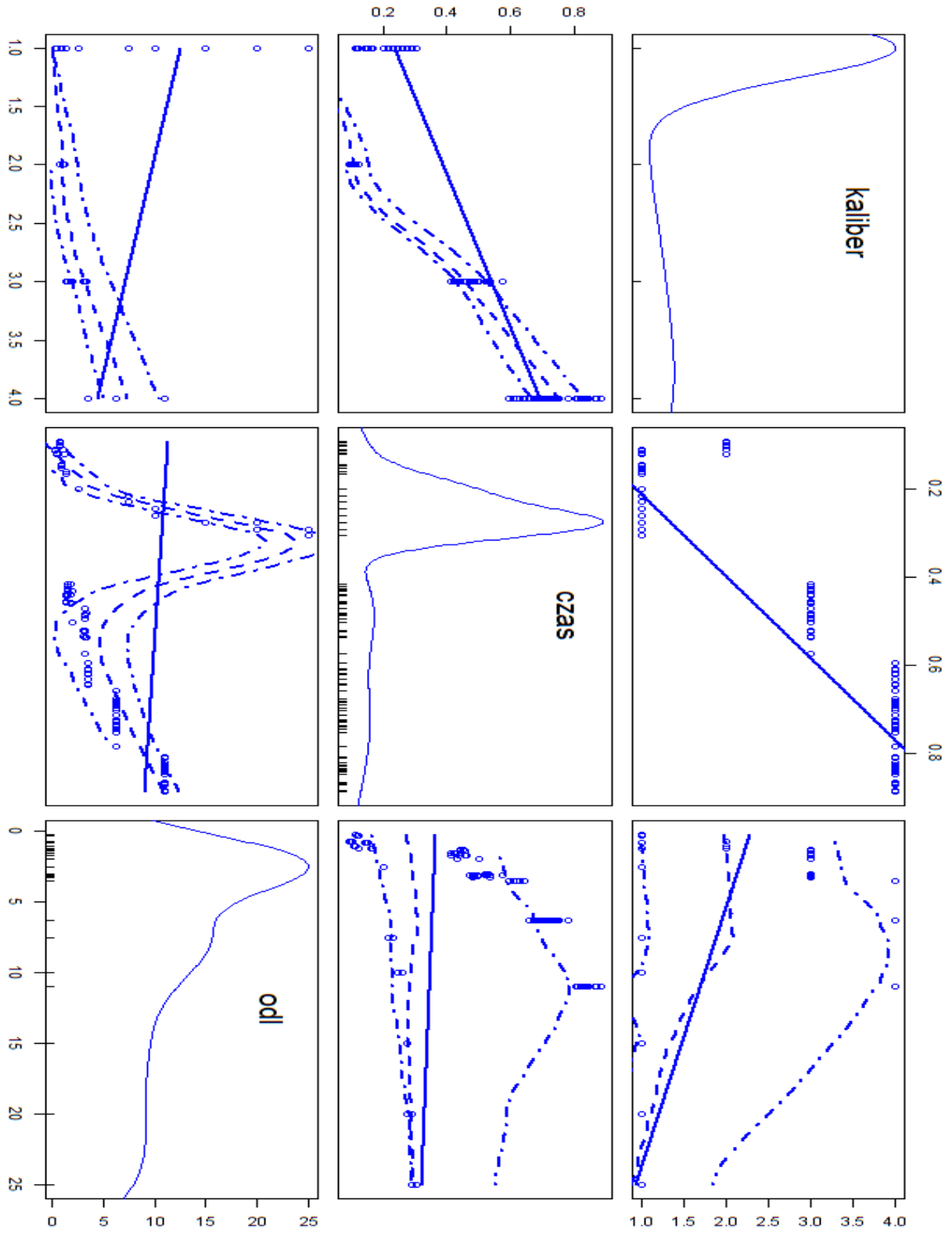
Wartości cząstkowe Wilksa dla poszczególnych zmiennych posiadają wartość p bliską zeru, zatem obie zmienne są istotne statystycznie. Niższa wartość statystyki dla zmiennej *czas*, przedstawiona w tabeli na rys. 26 oznacza, że ta zmienna osobno lepiej rozróżnia grupy niż zmienna *odl*.

Sprawdzenie spełnienia założeń LDA

Liniowa analiza dyskryminacyjna opiera się na założeniach: rozkładu normalnego zmiennych, ich braku współliniowości oraz równości macierzy kowariancji grup.

Przedstawiony fragment tabeli wyników analizy na rys. 26 posiada informacje o tolerancji, której wartość 0,382 << 1, co świadczy, że zmienne nie są całkowicie skorelowane ze sobą i mogą razem występować w modelu.

W celu poglądowego sprawdzenia normalności rozkładu zmiennych posłużono się korelogramem (rys. 27):



Rysunek 27. Korelogram zmiennych rozpatrywanego modelu klasyfikującego.

Z rys. 27 wynika, że jedynie zmienna *czas* nosi znamiona rozkładu normalnego, inne zmienne znacznie od niego odbiegają. Sprawdzono tą hipotezę testem Shapiro-Wilka (wyc. 2).

```
        Shapiro-Wilk normality test

data:  dane[, 1]
W = 0.62029, p-value < 2.2e-16

> shapiro.test(dane[,2])

        Shapiro-Wilk normality test

data:  dane[, 2]
W = 0.80244, p-value < 2.2e-16

> shapiro.test(dane[,3])

        Shapiro-Wilk normality test

data:  dane[, 3]
W = 0.88858, p-value = 2.097e-13
```

Wycinek 2. Test Shapiro-Wilka wszystkich zmiennych.

W przypadku wszystkich zmiennych odrzucona zostaje hipoteza zerowa, mówiąca o rozkładzie normalnym zmiennej.

W celu sprawdzenia równości macierzy kowariancji w każdej grupie, wykorzystano test M-Boxa (wyc. 3):

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  dane[, 2:3]
Chi-Sq (approx.) = 524.4, df = 9, p-value < 2.2e-16
```

Wycinek 3. Test M-Boxa równości macierzy kowariancji.

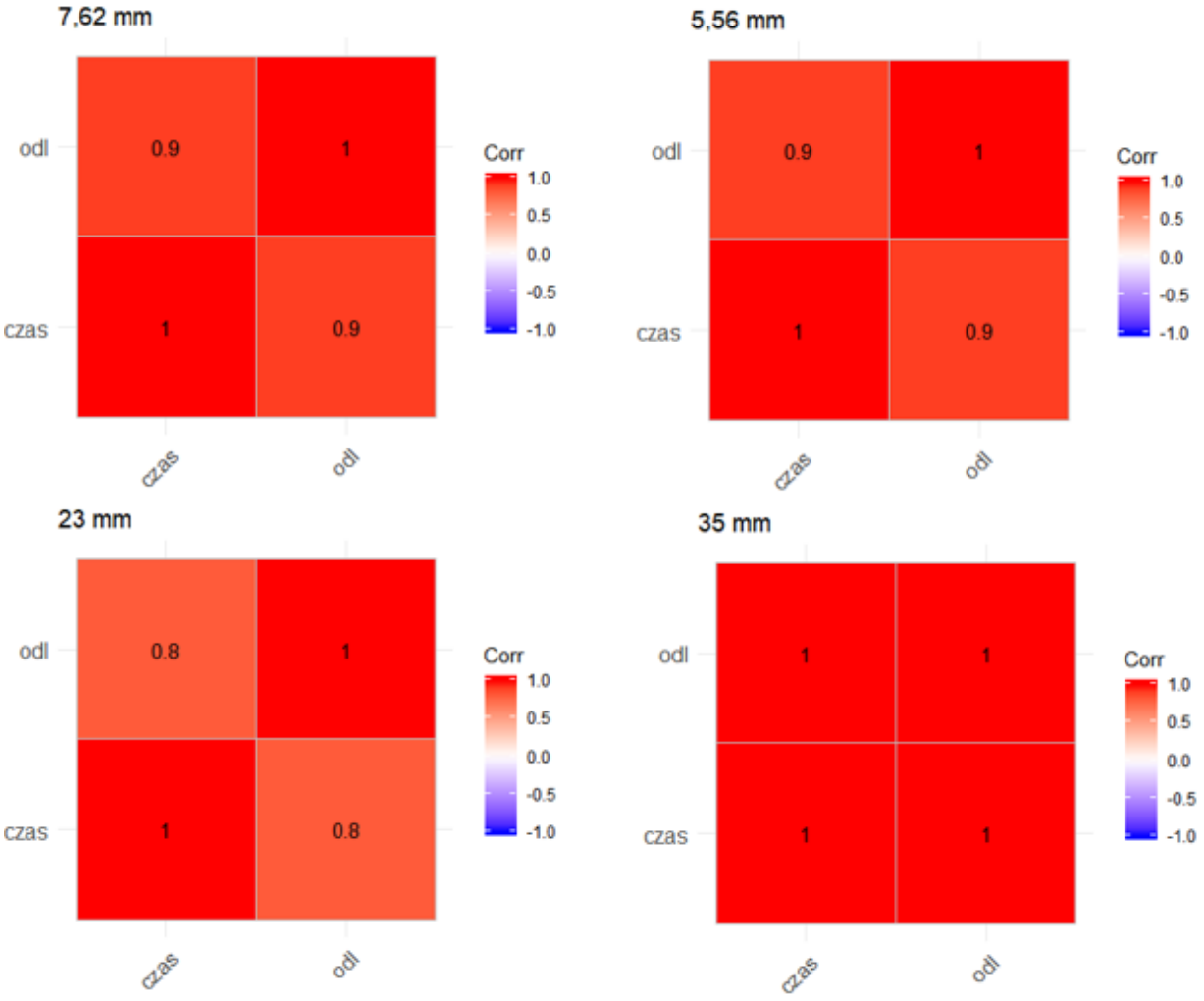
Hipoteza zerowa mówiąca o równości macierzy kowariancji nie została niestety utrzymana. W celu sprawdzenia natężenia odchyleń obliczono macierze kowariancji dla każdej grupy (wyc. 4).

```
> print(kowar1)
           czas        odl
czas 0.00211664  0.3485911
odl  0.34859110 68.8747546
> print(kowar2)
              czas         odl
czas 0.0001031556 0.00194400
odl  0.0019440000 0.04082667
> print(kowar3)
            czas        odl
czas 0.001899219 0.02835938
odl  0.028359377 0.63122776
> print(kowar4)
            czas       odl
czas 0.008125471 0.2517217
odl  0.251721746 8.5651696
```

Wycinek 4. Wartość macierzy kowariancji poszczególnych klas.

Macierze kowariancji dla grup 7,62 mm, 23 mm i 35 mm niewiele się od siebie różnią. Znacznie odstaje od nich macierz dla kalibru 5,56 mm; jest bardzo prawdopodobne, że to ona jest przyczyną negatywnego wyniku testu Shapiro-Wilka. Pewien ogląd na sytuację ujawnia

również wykres korelacji zmiennych w poszczególnych grupach, jako że korelacja jest liczona jako iloraz kowariancji do iloczynu błędów standardowych obu zmiennych. Wynik korelacji jest unormowaną i interpretowalną wartością kowariancji. Wykresy macierzy korelacji przedstawiono na rys. 28.

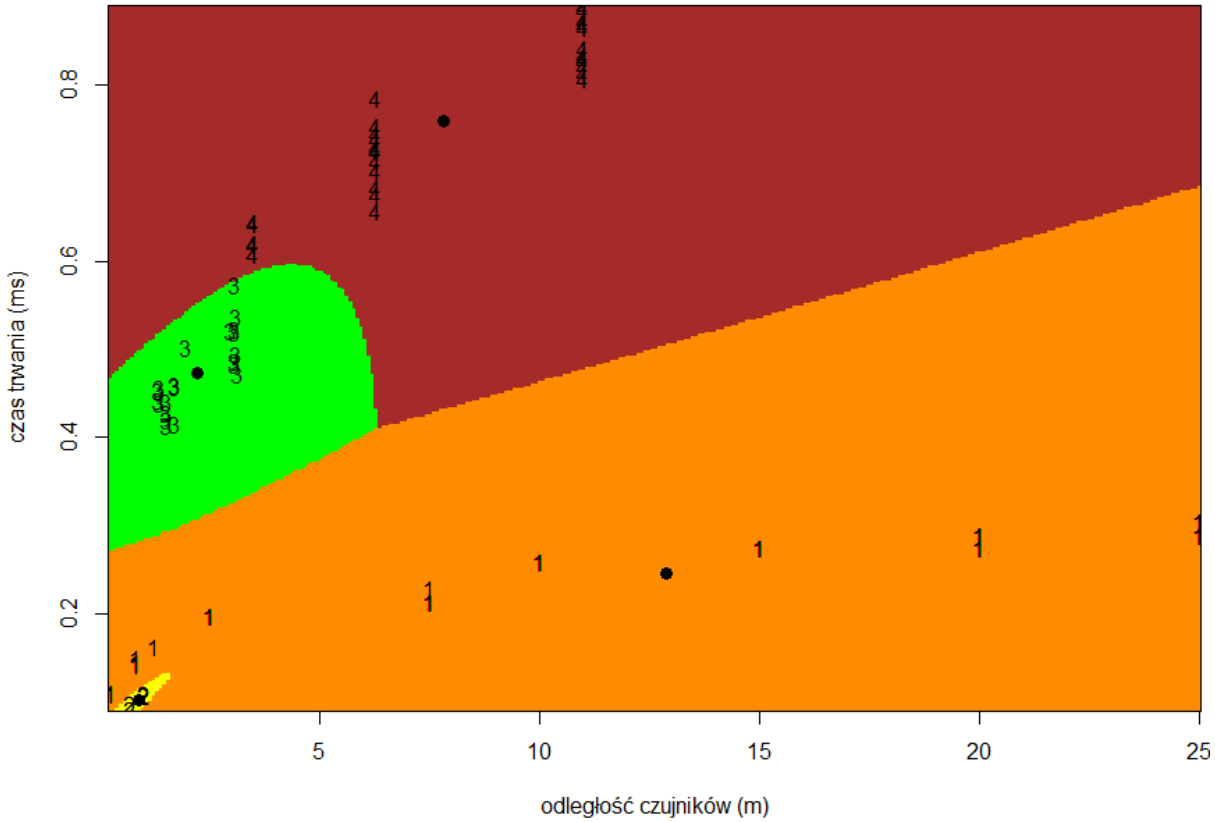


Rysunek 28. Wykresy macierzy korelacji.

Macierze korelacji w poszczególnych grupach przyjmują niemal identyczne wartości, co ostatecznie skłania do wniosku, że macierze kowariancji nie różnią się w sposób istotny od siebie.

4.2. Model QDA

W przypadku modelu QDA nie jest możliwe wyeksportowanie funkcji dyskryminujących z uwagi na ich skomplikowanie oraz nieinterpretowalność. Możliwe jest jednak zobrazowanie sklasyfikowania obserwacji do poszczególnych grup z zaznaczeniem granic funkcji dyskryminacyjnych (rys. 29):



Rysunek 29. Klasyfikacja przypadków i granice między grupami w modelu QDA.

Tabela klasyfikacyjna dla modelu QDA jest przedstawiona na wyc. 5.

	empiryczne			
teoretyczne	1	2	3	4
1	63	0	0	0
2	0	3	0	0
3	0	0	13	0
4	0	0	0	14

Wycinek 5. Tabela klasyfikacyjna QDA.

100% przypadków zostało poprawnie sklasyfikowanych. Widoczna jest wyższość elastycznych funkcji QDA względem liniowych LDA.

4.3. Model regresji logistycznej

Regresja logitowa jest metodą wyznaczania zmiennej zero-jedynkowej, dlatego podczas wyznaczania funkcji danego kalibru, wszystkie inne kalibry traktowano jako jedną zmienną zerową.

Oszacowano model regresji logistycznej dla kalibru 7,62 mm, otrzymując następujące wyniki (wyc. 6):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.7786     0.6746   4.119 3.81e-05 ***
odl           1.1692     0.5069   2.306  0.02109 *
czas        -22.2020     7.0938  -3.130  0.00175 **
      Likelihood ratio test

      Model 1: kaliber ~ odl + czas
      Model 2: kaliber ~ 1
        #Df   LogLik Df  Chisq Pr(>Chisq)
      1   3  -20.415
      2   1 -111.710 -2 182.59  < 2.2e-16 ***
```

Wycinek 6. Model logitowy 7,62 mm.

Końcowy model jest postaci:

$$Y_1 = 2{,}7786 + 1{,}1692 \cdot odl - 22{,}202 \cdot czas,$$

w którym prawdopodobieństwo, że pocisk jest kalibru 7,62 mm jest równe:

$$p_{1i} = \frac{e^{2{,}7786+1{,}1692\cdot odl-22{,}202\cdot czas}}{1 + e^{2{,}7786+1{,}1692\cdot odl-22{,}202\cdot czas}}$$

Ocena istotności

Według testu Walda poszczególnych zmiennych (ostatnia i przedostatnia kolumna wyc. 7) zarówno zmienna *odl*, jak i *czas* są statystycznie istotne dla poziomu istotności $\propto = 0{,}05$. Test istotności całego modelu (test ilorazu wiarygodności), także nakazuje odrzucenie hipotezy zerowej, że wszystkie zmienne są łącznie nieistotne.

Walidacja modelu

Tabela klasyfikacyjna jest przedstawiona na wyc. 7:

```
                Empiryczne
Teoretyczne  0  1
          0 27  0
          1  3 63
```

Wycinek 7. Tabela klasyfikacyjna regresji logitowej kalibru 7,62 mm.

Model poprawnie sklasyfikował 96,77% nowych przypadków, mając tendencję do klasyfikacji innych kalibrów jako 7,62 mm. Pomimo to, jakość klasyfikacji powyżej 90% jest uznawana jako zadowalająca.

Oszacowania modelu 5,56 mm (wyc. 8):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   46.223     29.051   1.591    0.112
odl            2.830      1.945   1.455    0.146
czas        -425.406    267.193  -1.592    0.111

Likelihood ratio test

Model 1: kaliber ~ odl + czas
Model 2: kaliber ~ 1
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   3  -3.0432
2   1 -29.7870 -2 53.488  2.428e-12 ***
```

Wycinek 8. Model regresji logitowej kalibru 5,56 mm.

Test istotności Walda (dwie ostatnie kolumny) wykazał, że żadna zmienna nie jest istotna statystycznie przy ogólnie przyjętych poziomach istotności. Oszacowania parametrów nie są więc stabilne. Przyczyną tego jest zapewne mała liczba obserwacji dla kalibru 5,56 mm.

Oszacowania modelu 23 mm (wyc. 9):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.855      1.077  -2.652 0.008004 **
odl           -2.619      0.755  -3.469 0.000523 ***
czas          20.622      5.166   3.991 6.57e-05 ***

Likelihood ratio test

Model 1: kaliber ~ odl + czas
Model 2: kaliber ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -18.320
2   1 -61.232 -2 85.823  < 2.2e-16 ***
```

Wycinek 9. Model regresji logitowej kalibru 23 mm.

Model:

$$Y_3 = -2{,}855 - 2{,}619 \cdot odl - 20{,}622 \cdot czas,$$

Prawdopodobieństwo:

$$p_{3i} = \frac{e^{-2{,}855-2{,}619 \cdot odl-20{,}622 \cdot czas}}{1 + e^{-2{,}855-2{,}619 \cdot odl-20{,}622 \cdot czas}}$$

Zarówno cały model, jak i poszczególne zmienne są istotne statystycznie.

Tabela klasyfikacyjna (wyc. 10):

```
                Empiryczne
Teoretyczne  0  1
          0 78  7
          1  2  6
```

Wycinek 10. Tabela klasyfikacyjna modelu 23 mm.

Model sklasyfikował poprawnie 90,33% przypadków.

Oszacowania dla kalibru 35 mm (wyc. 11):

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -884.02  144763.17  -0.006    0.995
odl            17.17    3090.32   0.006    0.996
czas         1416.81  231293.59   0.006    0.995
```

Wycinek 11. Oszacowanie regresji logitowej dla kalibru 35 mm.

Test Walda wykazał, że wszystkie zmienne występujące w modelu są nieistotne statystycznie. Oszacowane parametry nie są więc wiarygodne. Podobnie jak w przypadku kalibru 5,56 mm przyczyną może być dość niska ilość obserwacji.

4.4. Działanie algorytmu KNN

W tym przypadku nie był wyznaczany model, gdyż algorytm nie tworzy reprezentacji wiedzy o problemie na podstawie danych uczących, tylko znajduje rozwiązania w momencie pojawienia się wzorca testowego do klasyfikacji. Opracowany algorytm szukał najbardziej podobnej obserwacji pod względem zmiennych objaśniających i klasyfikował do grupy tej obserwacji nowy przypadek. Szukano metodą prób i błędów optymalnej liczby K sąsiadów. Okazała się nią liczba *K*=3, która także w innych badaniach często jest optymalna. Skuteczność klasyfikacji wyniosła 97,8% (wyc. 12). Podobnie jak w przypadku liniowej analizy dyskryminacyjnej, model miał problemy z właściwym sklasyfikowaniem dwóch pocisków kalibru 7,62 mm, przyjmując je jako pociski 5,56 mm.

```
                   EmpiryczneKNN
TeoretycznеKNN  1  2  3  4
             1 61  0  0  0
             2  2  3  0  0
             3  0  0 13  0
             4  0  0  0 14
```

Wycinek 12. Tabela klasyfikacyjna dla metody KNN.

W związku z charakterem działania algorytmu, nie jest możliwe wykreślenie granic decyzyjnych.

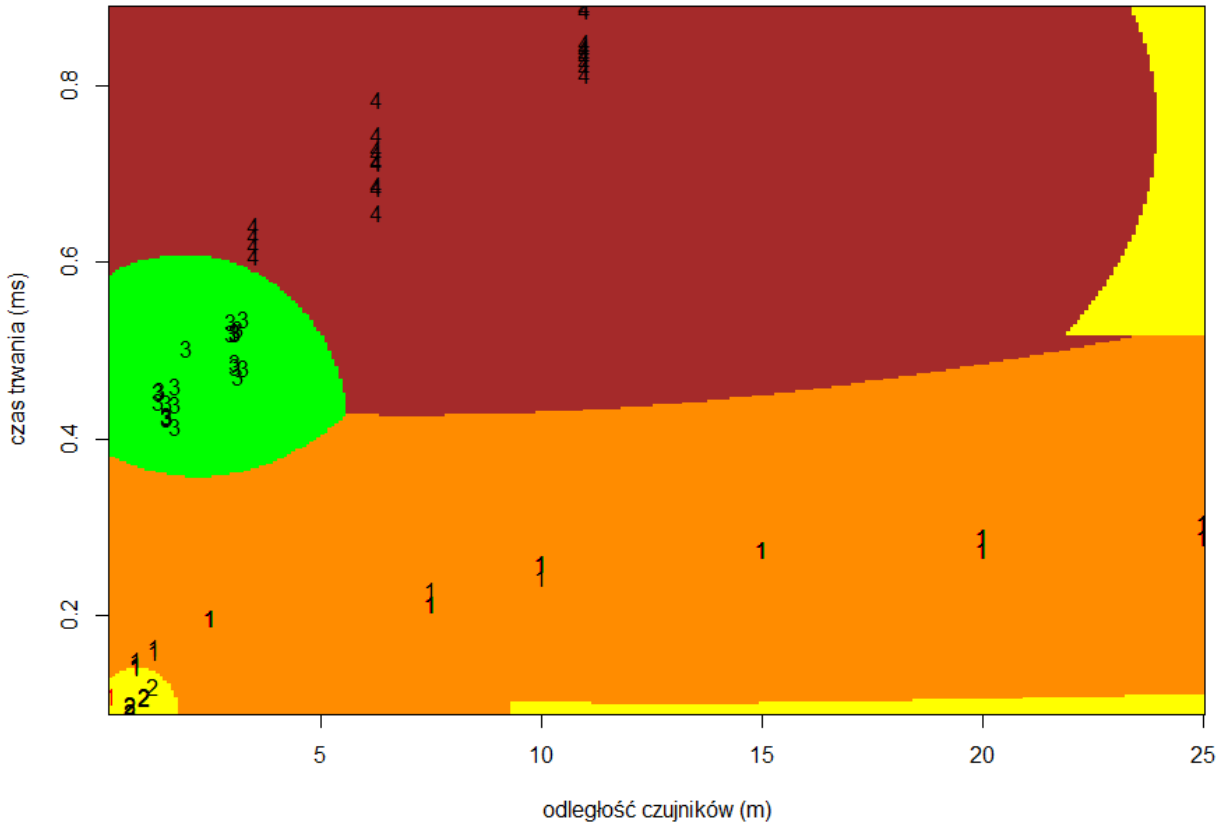
4.5. Działanie naiwnego klasyfikatora Bayesa

Klasyfikator oszacował prawdopodobieństwa *a posteriori* nowych przypadków na podstawie częstości i przyjmując rozkład Gaussa danych uczących. Pomimo faktycznej zależności zmiennych (klasyfikator przyjmuje, że zmienne są niezależne) otrzymano wysoką skuteczność klasyfikacji (wyc. 13):

	empiryczne			
oceny	1	2	3	4
1	59	0	0	0
2	2	5	0	0
3	0	0	11	0
4	0	0	0	16

Wycinek 13. Tabela klasyfikacyjna metody Bayesa.

Skuteczność klasyfikacji wyniosła 97,8%. Klasyfikator poprawnie rozróżnił kalibry 23 mm i 35 mm, jednak kalibry 5,56 mm i 7,62 mm okazały się zbyt trudne do rozróżnienia, przyjmując dwa pociski 7,62 mm jako 5,56 mm. Granice decyzyjne przyjmują kształt liniowo-paraboliczny (rys. 30).



Rysunek 30. Obszary decyzyjne w klasyfikatorze Bayesa.

Granice decyzyjne na rys. 30 są nienaturalne i niemożliwe do osiągnięcia, szczególnie dla kalibru 5,56 mm. Jest to przesłanką do stwierdzenia, że klasyfikator Bayesa nie jest poprawną metodą do klasyfikacji tego typu danych.

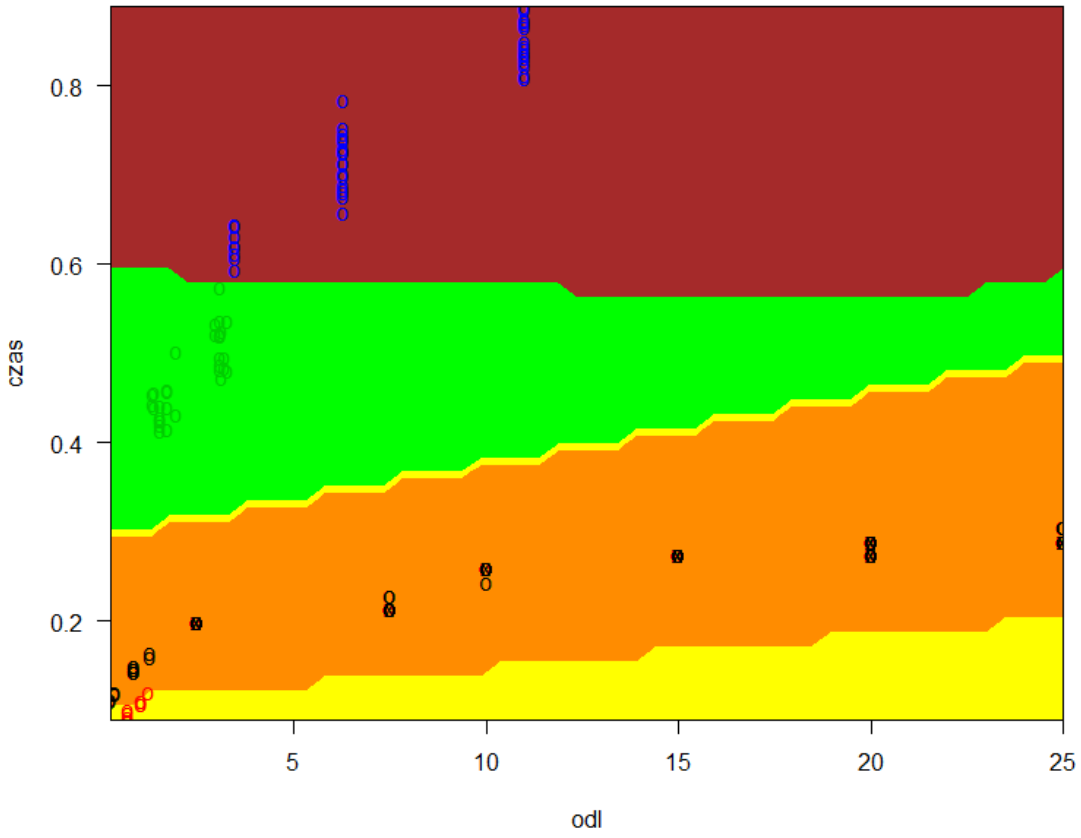
4.6. Wykorzystanie metody SVM

Zastosowano metodę wektorów nośnych, przekształcając dane w przestrzeń liniową oraz radialną. Empirycznie znaleziono optymalną wartość kosztu kary wynoszącą 7. Dla tej wartości osiągnięto stuprocentową skuteczność klasyfikacji, zarówno dla jądra liniowego, jak i radialnego (wyc. 14). Domyślna wartość kosztu kary w pakietach statystycznych równa 1, spowodowałaby pogorszenie skuteczności klasyfikacji.

	oceny			
empiryczne	1	2	3	4
1	61	0	0	0
2	0	5	0	0
3	0	0	11	0
4	0	0	0	16

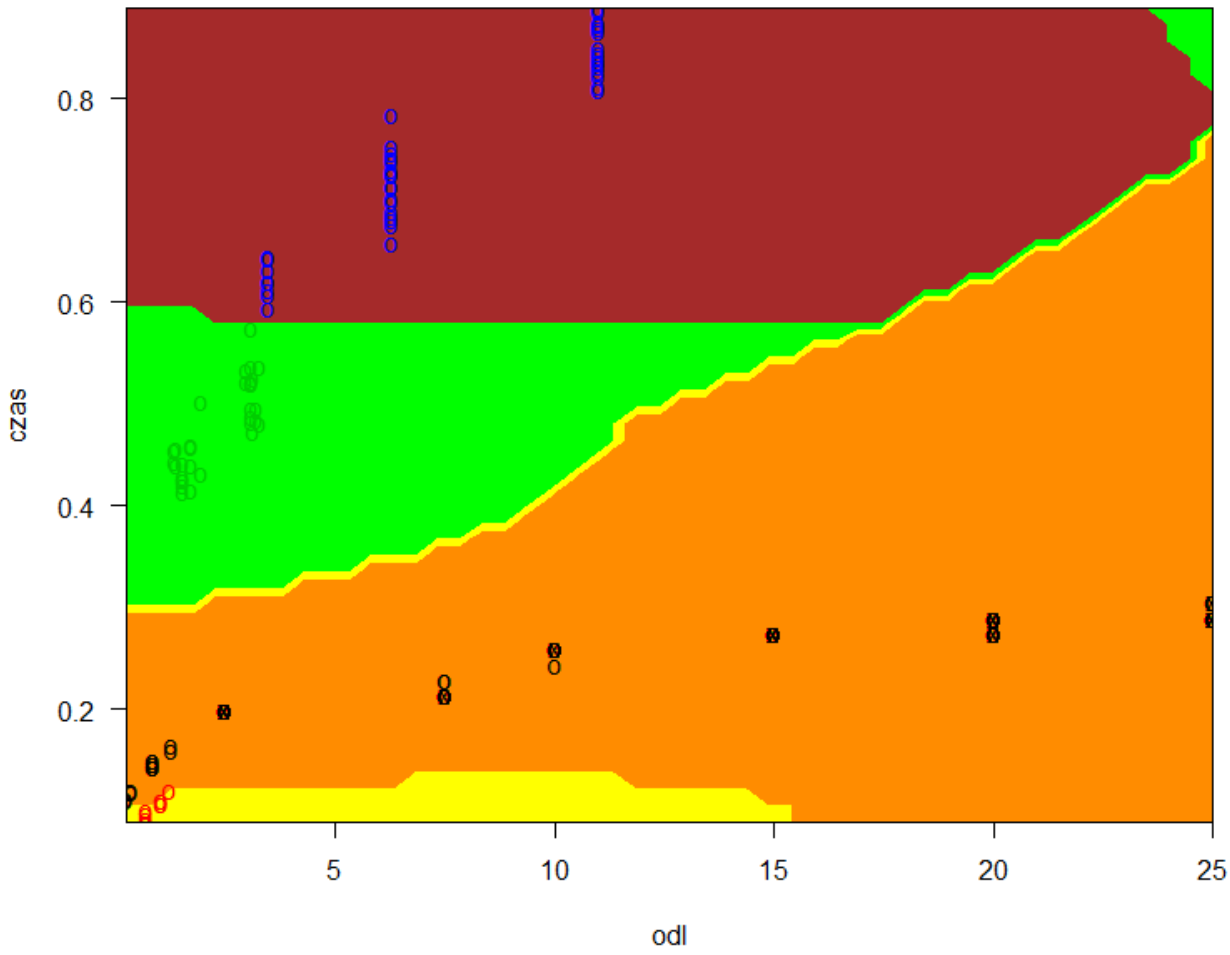
Wycinek 14. Wyniki klasyfikacji SVM.

Wyznaczone obszary decyzyjne dla jądra liniowego i radialnego, przedstawiają odpowiednio rys. 31 oraz rys. 32.



Rysunek 31. Obszary decyzyjne- SVM liniowe.

Wyrysowane obszary decyzyjne na rys. 31 wydają się realistyczne i są podobne do obszarów pochodzących z liniowej analizy dyskryminacyjnej (co zrozumiałe, ponieważ przekształcono dane do przestrzeni liniowej), jednak w tym przypadku osiągnięto stuprocentową skuteczność klasyfikacji. Jedynie wyrysowana granica między kalibrami 23 i 35 [mm] może być niezgodna z rzeczywistością dla większych odległości (powodem jest brak chociażby jednej obserwacji dla kalibru 23 mm dla dłuższego dystansu).

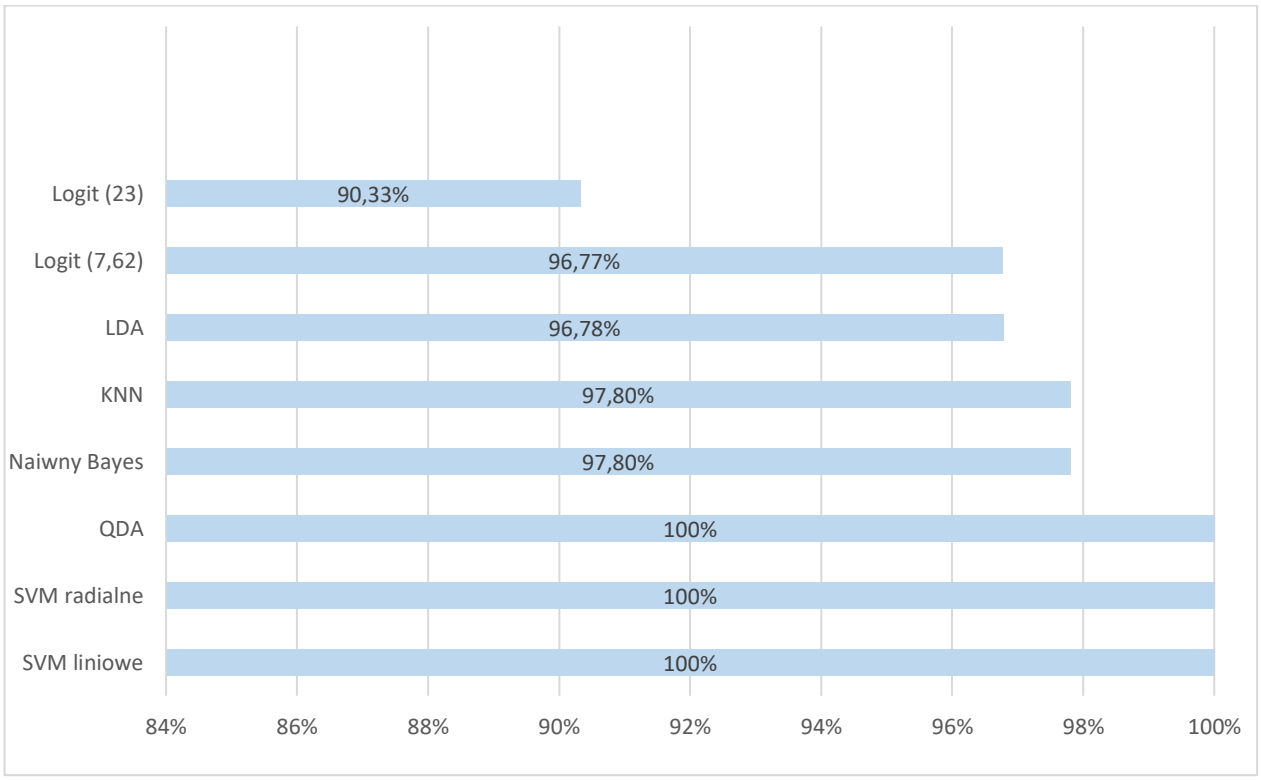


Rysunek 32. Klasyfikacja przypadków - SVM radialne.

Nierealistyczne obszary decyzyjne z rys. 32 (szczególnie dla kalibru 7,62 mm), świadczą że wykorzystanie jądra radialnego nie jest wskazane dla tego typu zbioru danych.

4.7. Wnioski końcowe

Porównanie skuteczności metod klasyfikacji zostało przedstawione na rys. 33:



Rysunek 33. Porównanie skuteczności metod klasyfikacji.

Pierwszym narzucającym się wnioskiem jest bardzo dobra skuteczność klasyfikacji dla każdej metody. Świadczy to o tym, że pociski można skutecznie rozróżniać, stosując odległość czujników od trajektorii lotu oraz czas trwania fali N jako predyktory. Największy problem sprawiało rozróżnienie kalibrów 5,56 mm i 7,62 mm, jednak metody QDA i SVM poradziły sobie z tym bezbłędnie.

W przypadku kalibrów 5,56 mm oraz 35 mm, wyznaczenie modeli regresji logitowej nie było możliwe z uwagi na nieistotność statystyczną zmiennych. Potwierdziło się przypuszczenie, że regresja logistyczna ma spore problemy z niestabilnością parametrów, gdy klasy są w dużej mierze znacznie rozdzielone od siebie. Negatywny wpływ miała także dość niska ilość obserwacji, na co regresja logitowa również jest podatna. Budowa modelu klasyfikującego pociski wymaga jednak skutecznego rozróżniania pocisków nawet w przypadku małej ilości obserwacji. Pozyskiwanie danych o fali N wymaga zorganizowanych przedsięwzięć i jest kosztowne szczególnie w przypadku amunicji większych kalibrów. Z tych względów regresja logistyczna nie może być optymalnym klasyfikatorem.

W celu wyboru najlepszej metody klasyfikacji należy zwrócić uwagę na ułożenie się danych. Zarówno wykres wartości teoretycznych (rys. 9), jak i empirycznych (rys. 24) czasu trwania fali N w funkcji odległości wskazują, że obserwacje w klasach układają się w lekką parabolę. To indukuje, że optymalna jest kwadratowa analiza dyskryminacyjna (dodatkowo ta metoda pozwoliła na uzyskanie stuprocentowej skuteczności rozróżniania). Tak jednak nie jest. Przyglądając się wykresowi z rys. 29 można zauważyć, że funkcje dyskryminujące wyznaczyły bardzo małe obszary decyzyjne dla kalibrów 5,56 mm i 23 mm. W rzeczywistości obszary te powinny być w przybliżeniu równe dla wszystkich klas. Przyczyną wyznaczenia takich obszarów decyzyjnych, były znacząco różne wartości maksymalne i minimalne zmiennych (np. mały rozrzut dla kalibru 5,56 mm i duży dla 7,62 mm), to z kolei jest oznaką braku normalności rozkładu zmiennych. W rzeczywistych warunkach, trudno jest oczekiwać by zmienne: czas trwania fali N oraz odległość do toru lotu pocisku, układały się w rozkład normalny. Skuteczność kwadratowej analizy dyskryminacyjnej jest bardzo zależna od prawdziwości gaussowskiego rozkładu [47], którego zmienne w modelu nie posiadają, zatem i ta metoda nie jest optymalna.

Także naiwny klasyfikator Bayesa w celu znalezienia prawdopodobieństwa, przyjmuje rozkład gaussowski, zatem i ta metoda nie jest odpowiednia do klasyfikacji pocisków, co jest widoczne w bardzo różnych obszarach decyzyjnych na rys. 30.

Liniowa analiza dyskryminacyjna również wymaga rozkładu normalnego, ale jest bardziej odporna na niespełnienie tego założenia, co zostało wspomniane w rozdziale drugim. Odchylenia od rozkładu Gaussa w rozpatrywanym w modelu nie są tak duże, by znacząco wpłynąć na niedokładność parametrów w przypadku tej metody. Ponadto macierze kowariancji w grupach są zbliżone do równych, co jest spełnieniem kolejnego założenia LDA. Liniowa analiza dyskryminacyjna nigdy nie będzie metodą w stu procentach skuteczną, gdyż obszary decyzyjne grup zawsze będą w niewielkim stopniu się wzajemnie przenikać, szczególnie w przypadku kalibrów o małej różnicy w wielkości (np. 5,56 i 7,62 mm). Poszczególne grupy w modelu klasyfikującym pociski są jednak na tyle oddzielne (rys. 25), że LDA jest metodą, która może z powodzeniem zostać wykorzystana do rozróżniania pocisków. Potwierdza to wysoka skuteczność klasyfikacji 96,78%.

Metoda KNN również osiągnęła bardzo wysoki rezultat i może być z powodzeniem stosowana do klasyfikacji wielu grup pocisków. Algorytm ten jest szczególnie skuteczny,

gdy posiada się duży zbiór danych uczących. Wadą tej metody jest nieskuteczne rozróżnianie małych kalibrów i brak tworzenia reprezentacji o problemie.

Ostatnia analizowana metoda wektorów nośnych (SVM), okazała się najskuteczniejsza wśród metod uczenia maszynowego. Należy odrzucić opcję przekształcania danych do przestrzeni radialnej, gdyż powoduje to wyznaczenie nienaturalnych obszarów decyzyjnych. Jednakże przekształcenie danych do przestrzeni liniowej, umożliwiło wyznaczenie obszarów decyzyjnych zbliżonych do liniowych (jak w przypadku LDA) przy zachowaniu stuprocentowej skuteczności klasyfikacji.

Podsumowanie

Celem pracy było wybranie odpowiedniego klasyfikatora do rozróżniania pocisków na podstawie parametrów charakterystycznych fali N. Poprzez analizę działania wybranych metod statystycznych, udowodniono hipotezę, że na podstawie parametrów fali N pociski mogą być skutecznie identyfikowane.

W świetle omawianej teorii w rozdziale pierwszym, zwrócono uwagę na dwa aspekty, które świadczą o tym, że pociski można rozróżniać na podstawie parametrów fali N. Pierwszym aspektem jest to, że rozkład parametrów fali N można modelować za pomocą liniowych równań Whithama i na tej podstawie decydować, jakie zmienne należałoby wybrać jako predyktory do budowy modeli klasyfikacyjnych pocisków. Drugim istotnym aspektem jest fakt, że czas trwania fali N i jej szczytowa wartość amplitudy są parametrami w pełni identyfikującymi źródło zarejestrowanego zaburzenia ciśnienia, jakim jest pocisk danego kalibru będący w ruchu naddźwiękowym. Warunki środowiskowe (np. absorbcja powietrza) mogą mieć wpływ na pomiar wartości parametrów generowanych przez pocisk zaburzeń ciśnieniowych. Zatem w celu budowy modelu o najwyższej wiarygodności, należy koniecznie opierać się na dużej ilości danych empirycznych (statystycznych).

W wyniku dalszej analizy sformułowano wniosek, że ciśnienie szczytowe fali N nie jest dobrym dyskryminatorem, ponieważ jego pomiar determinuje zastosowanie mikrofonów o takiej samej czułości, w celu pozyskania porównywalnych wyników, co jest niepraktyczne w świetle dalszych zastosowań. Zmienną, którą wybrano jako najlepiej rozróżniającą pociski jest czas trwania fali N. Na zmienną tą ma wpływ odległość mikrofonów i prędkość pocisku, więc również one (lub przynajmniej jedno z nich) powinny być uwzględnione w modelu klasyfikacyjnym. Zmiennej *prędkość pocisku* nie uwzględniono w pracy ze względu na brak danych. Nie wiadomo czy byłaby ona istotna statystycznie. Końcowe modele zawierają zatem następujące zmienne: *czas* trwania fali N oraz *odległość* mikrofonów od trajektorii lotu pocisków.

Zebrano dostępne dane z czterech strzelań dla kalibrów 5,56; 7,62; 23 oraz 35 [mm]. W każdym przypadku obliczono czas trwania fali N na podstawie wyeksportowanych plików PicoScope, tekstowych lub zmiennoprzecinkowych, co odbywało się za pomocą napisanego programu w języku Python oraz programu MS Excel. W niektórych analizowanych przypadkach nie była podana wprost wartość odległości toru lotu pocisku od mikrofonów

pomiarowych i obliczano ją na podstawie geometrii rozłożenia stanowiska pomiarowego i znajomości położenia punktu trafienia pocisku w tarczę. Łącznie zebrano dane o 278 przebiegach fali N.

Spośród stosowanych obecnie metod klasyfikacji, przeanalizowano wybrane metody mogące być użyteczne do rozróżniania pocisków na podstawie parametrów fali N. Zastosowano liniową i kwadratową analizę dyskryminacyjną, regresję logistyczną oraz metody uczenia maszynowego: najbliższych sąsiadów (KNN), wektorów nośnych (SVM) oraz naiwny klasyfikator Bayesa. Analiza otrzymanych wyników wykazała, że regresja logistyczna ma problemy za stabilnością parametrów oraz jest podatna na niską ilość obserwacji, więc nie jest odpowiednią metodą w przypadku klasyfikacji pocisków. Odrzucono także kwadratową analizę dyskryminacyjną oraz naiwny klasyfikator Bayesa, ponieważ te metody okazały się być bardzo wrażliwe na odchylenia od rozkładu normalnego zmiennych, które wystąpiły w przypadku parametrów fali N.

Odpowiednimi metodami służącymi do klasyfikacji pocisków okazały się być: liniowa analiza dyskryminacyjna oraz metody K-najbliższych sąsiadów i wektorów nośnych. Skuteczność klasyfikacji nowych danych dla tych metod przekroczyła 96%; w przypadku SVM osiągnięto 100%. W metodach LDA i KNN pojawiało się błędne sklasyfikowanie między kalibrami 7,62 mm i 5,56 mm, prawdopodobnie z powodu niewielkich różnic w wymiarach obu typów pocisków, co wpływa na wartość generowanych podczas lotu pocisku parametrów fali N.

Wybór odpowiedniej metody zależy od efektów jakie chcemy uzyskać. Jeśli zależy nam na skonstruowaniu w pełni interpretowalnego modelu przy zachowaniu wysokiej skuteczności klasyfikacji, wtedy najlepszą metodą jest liniowa analiza dyskryminacyjna. Trzeba wtedy jednak liczyć się ze sporadycznymi błędami klasyfikacji między pociskami o niewiele różniących się wymiarach. Jeśli zależy nam na pełnej skuteczności identyfikacji, to przy użyciu metody wektorów nośnych jest to możliwe.

Wyniki pracy wskazały najskuteczniejszy klasyfikator, jaki może zostać użyty w systemach identyfikacji pocisków na podstawie fali N. Ponadto skonstruowany model LDA dla kalibrów 5,56; 7,62; 23; 35 [mm] jest użyteczny i może zostać wdrożony w systemach identyfikacji uzbrojenia.

Bibliografia

[1] Danicki, E. (2006). *The shock wave-based acoustic sniper localization*. Nonlinear Analysis: Theory, Methods & Applications, 65(5), 956–962. DOI:10.1016/j. na 2005.07.043.

[2] Volgyesi, P., Balogh, G., Nadas, A., Nash, C. B., & Ledeczi, A. (2007). *Shooter localization and weapon classification with soldier-wearable networked sensors*. Proceedings of the 5th International Conference on Mobile Systems, Applications and Services - MobiSys '07. DOI:10.1145/1247660.1247676.

[3] Eckert, J., Carpenter, M., Hartfield, R., & Cervantes, N. (2020). *Classification of Intermediate Range Missiles During Launch. AIAA Scitech 2020 Forum.* DOI:10.2514/6.2020-1852.

[4] Singh U. K., Padmanabhan V. (2013), *Training by ART-2 and Classification of Ballistic Missiles Using Hidden Markov Model.* Pattern Recognition and Machine Intelligence, Springer, 108-115.

[5] Singh U. K., Padmanabhan V., Agarwal A. (2013), *Novel Method For Training and Classification of Ballistic and Quasi-Ballistic Missiles in Real-Time,* Proceedings of International Joint Conference on Neural Networks, Dallas, Texas, USA, August 4-9, 2933-2940.

[6] Damarla T., Kaplan L., Whipps G. (2010), Sniper Localization Using Acoustic Asychronous Sensors, IEEE Sensors Journal, Vol. 10, No. 9, 1469-1473.

[7] Songyun, G., Yi, W., & Yaohui, Z. (2011), *The Sound Recognition of Artillery Projectile Shape Based on Least Squares Fuzzy Support Vector Machine. 2011 International Conference on Intelligence Science and Information Engineering.*

[8] Muck R. (1973), *Study of methods which predict supersonic flow fields form body geometry, distance and Mach number,* Va 23665, NASA TN 0-7387, 3-13.

[9] Cramer O.(1993), *The variation of the specific heat ratio and the speed of sound in the air with temperature, pressure, humidity, and CO2 concentration*, „Journal of the Acoustical Society of America", Vol. 93, No. 5, 2510-2516.

[10] Loucks R., Bradford B., Davis S., Moss L., Pham T., Fong M. (1995), *Method of Identifying Supersonic Projectiles Using Acoustic Signatures,* Army Research Laboratory, ARL-TR-859, 3-80.

[11] Pietrasieński J., Rodzik D., Warchulski J., Warchulski M, (2006), *Analiza częstotliwościowa zaburzeń ośrodka wywołanych lotem pocisku*, Biuletyn WAT, Vol. LVI, nr specjalny (1); s 433-443.

[12] Maher C. (2007), *Acoustical Characterization of Gunshots*, „2007 IEEE Workshop on Signal Processing Applications for Public Security and Forensics", 109-114.

[13] Raschka S., Mirjalili V.(2017), *Python. Uczenie maszynowe. Wydanie drugie*, Helion, 109-113.

[14] Aczel D.A. (2000), *Statystyka w zarządzaniu*, PWN, 883.

[15] Dobosz M. (2004), *Wspomagana komputerowo statystyczna analiza wyników badań*, AOW EXIT, 90.

[16] Hadasik D. (1998), *Upadłość przedsiębiorstw w Polsce i metody jej prognozowania*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, 117.

[17] Lachenbruch P. A. (1975), *Discriminant analysis*, Hafner Press.

[18] Radkiewicz P. (2010), *Analiza dyskryminacyjna. Podstawowe założenia i zastosowania w badaniach społecznych*, Psychologia Społeczna tom 5, 142-161.

[19] Friendly M., Sigal M. (2014), *Visualizing Tests for Equality of Covariance Matrices*, „National Sciences and Engineering Research Council of Canada" [RGPIN-03772-2014].

[20] Gruszczyński M., Kuszewski T., Podgórska M. (2009), *Ekonometria i badania operacyjne*, PWN, 57-59.

[21] Kasjaniuk M. (2006), *Zastosowanie analizy dyskryminacyjnej do modelowania i prognozowania kondycji przedsiębiorstw*, „Barometr regionalny 6/2006", 95-100.

[22] Wojtatowicz T. (1998), *Metody analizy danych doświadczalnych. Wybrane zagadnienia*, Politechnika Łódzka, 34-35.

[23] Maddala E. (2006), *Ekonometria*, PWN, 371.

[24] Gruszczyński M. (red.) (2010), *Modele i metody analizy danych indywidualnych*, Wolters Kluwer, 62-64.

[25] Podgórska M.(red.) (2004), *Ekonometria*, Oficyna Wydawnicza SGH, 33-34.

[26] Park H. (2013), *An intruduction to logistic regression: from basic concepts to interpretation with particular to nursing domain*, "J Korean Acad Nurs Vol.43 No.2 April 2013", 158-159.

[27] Hosmer D., Lemeshow S. (2000), *Applied Logistic Regression. Second Edition*, Wiley-Interscience, 156-158.

[28] James G. (2017), *An Introduction to Statistical Learning*, Springer, 39-42.

[29] Tymków P. (2008), *Klasyfikacja obrazów rastrowych z wykorzystaniem sztucznych sieci neuronowych i statystycznych metod klasyfikacji*, Wydawnictwo Uniwersytetu Przyrodniczego we Wrocławiu, 21-24.

[30] Marcinkowska-Ochtyra A. (2016), *Ocena przydatności obrazów hiperspektralnych APEX oraz maszyn wektorów nośnych (SVM) do klasyfikacji roślinności subalpejskiej i alpejskiej Karkonoszy*. Rozprawa doktorska, Uniwersytet Warszawski, 35-42.

[31] Rodzik D. (2011), Wyniki badań eksperymentalnych realizowanych w ramach projektu badawczego promotorskiego nr O N501 165038 pt. *Opracowanie metody określania współrzędnych toru pocisku na podstawie parametrów fali typu N*, WAT, (niepublikowane).

[32] Pietrasieński J., Rodzik D. (2006), Sprawozdanie z pracy badawczej własnej nr PBW GR 577 nt. *Badanie rozchodzenia się zaburzeń ośrodka wywołanych ruchem pocisku*, WAT.

[33] Miernik J. (2019), *Badanie czujników i wyznaczenie parametrów fali N w funkcji odległości propagacji*, sprawozdanie z badań na strzelnicy poligonowej WITU w dniu 14.12.2018 r., WAT (niepublikowane).

[34] Rodzik D. (2018), *Badania symulacyjne i doświadczalne charakterystyk zaburzeń ciśnieniowych pochodzących od ruchu 35 mm pocisku TP-T na potrzeby budowy akustycznego systemu oceny strzelań wspomagającego badania kwalifikacyjne 35 mm okrętowego systemu uzbrojenia*, sprawozdanie z badań realizowanych w ramach projektu nr O ROB 0046 03 001 pt. „35 mm automatyczna armata morska KDA z zabudowanym na okręcie system systemem kierowania ogniem wykorzystującym Zintegrowaną Głowicą Śledzącą ZGS-158 wykonaną w wersji morskiej wraz ze stanowiskiem kierowania ogniem", WAT, (niepublikowane).

[35] Thoma K., Hornemann U., Sauer M., Schneider E. (2005), *Shock waves phenomenology, experimental and numerical simulation*. „Meteoritics & Planetary Science 40", Nr 9/10, 1283–1298.

[36] Maher C. (2006), *Modelling and signal processing of acoustic gunshot recordings*, „IEEE Signal Processing Society 12th DSP Workshop", 10, 257-261.

[37] Maher C. (2018), *Principles of Forensic Audio Analysis*, „Springer International Publishing".

[38] Whitham G. B. (1952), *The Flow Patterns of a Supersonic Projectile*, „Communications on Pure and Applied Mathematics", Vol. V, 301-348.

[39] Whitham G. B. (1950), *The Behaviour of Supersonic Flow Past a Body of Revolution*, „Far From the Axis, Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences", Vol. 201, No. 1064, 89-109.

[40] Miller, David S.; Morris, Odell A, Harry W. (1971), *Wind-Tunnel Investigation of Sonic-Boom Characteristics of Two Simple Wing Models at Mach Numbers From 2.3 to 4.63*. NASA TN D-6201.

Źródła internetowe:

[41] https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-inr/

[42] https://mlweb.loria.fr/book/en/lda.html

[43] https://scikit-learn.org/stable/modules/lda_qda.html

[44] https://www.mghassany.com/MLcourse/discriminant-analysis.html#quadratic-discriminant-analysis-qda

[45] http://www.statystyka.c0.pl/blog/LDA.html#jak_to_działa,_czyli_łyk_teorii

[46] Harańczyk G., *Krzywe ROC, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia.* media.statsoft.pl/_old_dnn/downloads/krzywe_roc_czyli_ocena_jakosci

[47] http://www.mif.pg.gda.pl/homepages/kdz/StatystykaII/Klasyfikacja.pdf

[48] https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

Załączniki

Załącznik 1. Kod źródłowy programu R wykorzystany w pracy do klasyfikacji

```
### wczytanie danych
dane=read.csv("C:\\Users\\user\\Desktop\\bazadanych.csv", header=TRUE, sep=';')

### wczytanie bibliotek
library(MASS)
library(car)
library(ggcorrplot)
library(caret)
library(VIF)
library(class)
library(MASS)
library(klaR)
library(biotools)
library(SpatialEpi)
library(lmtest)
library(e1071)
### rzut oka na dane
scatterplotMatrix(dane)


### losowanie indeksów wierszy zbioru uczącego
zbior.uczacy = sample(1:nrow(dane), nrow(dane)/1.5, F)
zbior.uczacy


### naiwny bayes

model=naiveBayes(factor(kaliber)~odl+czas, data = dane, subset = zbior.uczacy)
oceny=predict(model, newdata = dane[-zbior.uczacy,2:3], type = c("class"))
ct=table(oceny, empiryczne = dane[-zbior.uczacy,1])
ct
diag=(prop.table(ct,1))
sum(diag(prop.table(ct)))
 partimat(factor(dane$kaliber)~dane$czas+dane$odl, data = dane, subset=zbior.uczacy
         ,method = "naiveBayes", plot.matrix = FALSE,
         imageplot=TRUE, image.colors=c('cyan', 'deepskyblue','blue','darkblue'),
         col.correct='black', col.wrong='red', main="Klasyfikacja przypadków",
         prec=300, name=c('czas trwania (ms)','odległość czujników (m)'),
         print.err=0)


### metoda wektorów nośnych

svmfit=svm(factor(kaliber)~czas+odl, data = dane, subset = zbior.uczacy, kernel='linear',
           cost=10)
svmfit=svm(factor(kaliber)~czas+odl, data = dane, subset = zbior.uczacy, kernel='radial',
           cost=10)
print(svmfit)
summary(svmfit)
oceny=predict(svmfit, newdata = dane[-zbior.uczacy,2:3], type = c("class"))
oceny
ct=table(empiryczne = dane[-zbior.uczacy,1], oceny)
ct
diag=(prop.table(ct,1))
sum(diag(prop.table(ct)))
plot(svmfit, dane, subset = -zbior.uczacy, col=c('cyan', 'deepskyblue','blue','darkblue'))
```

```
57
58  ### liniowa analiza dyskryminacyjna
59  klasyfikatorLDA = lda(dane[,2:3], grouping = dane[,1], subset=zbior.uczacy)
60  print(klasyfikatorLDA)
61
62  ### wyświetlenie wykresów funkcji klasyfikujących
63  plot(klasyfikatorLDA, col='blue', cex=1.1)
64  title ("Wartości funkcji kanonicznych")
65  dane$kaliber=as.factor(dane$kaliber)
66  partimat(factor(dane$kaliber)~dane$czas+dane$odl, data = dane, subset=zbior.uczacy
67           ,method = "sknn", plot.matrix = FALSE,
68           imageplot=TRUE, image.colors=c('cyan', 'deepskyblue','blue','darkblue'),
69           col.correct='black', col.wrong='red', main="Klasyfikacja przypadków",
70           prec=300, name=c('czas trwania (ms)','odległość czujników (m)'),
71           print.err=0)
72
73  ### tabela klasyfikacyjna po użyciu zbioru testowego
74  oceny = predict(klasyfikatorLDA, newdata=dane[-zbior.uczacy,2:3])
75  ct=table(teoretyczne=oceny$class, empiryczne = dane[-zbior.uczacy,1])
76  ct
77
78  ### wyświetlenie procentu poprawnej klasyfikacji
79  diag=(prop.table(ct,1))
80  sum(diag(prop.table(ct)))
81
82  ### badanie współliniowości zmiennych
83  vif(dane$czas,dane$odl)
84
85  ### wyświetlenie rozkładu zmiennych
86  hist(dane[,2])
87  hist(dane[,3])
```

```
88
89  ### test rozkładu normalnego
90  shapiro.test(dane[,1])
91  shapiro.test(dane[,2])
92  shapiro.test(dane[,3])
93
94  ### obliczenie macierzy kowariancji
95  dane1=dane[1:194,2:3]
96  dane2=dane[195:204,2:3]
97  dane3=dane[205:236,2:3]
98  dane4=dane[237:278,2:3]
99
100 kowar1=cov(dane1)
101 kowar2=cov(dane2)
102 kowar3=cov(dane3)
103 kowar4=cov(dane4)
104
105 print(kowar1)
106 print(kowar2)
107 print(kowar3)
108 print(kowar4)
109
110 ### test M-Boxa
111 res=boxM(dane[,2:3],dane[,1])
112 res
113
114 ### wyświetlenie macierzy korelacji
115 korelacja1=round(cor(dane1),1)
116 ggcorrplot(korelacja1,lab = TRUE, title = "7,62 mm",col=c('red','darkblue','cyan'))
117 korelacja2=round(cor(dane2),1)
118 ggcorrplot(korelacja2,lab = TRUE, title = "5,56 mm", col=c('red','darkblue','cyan'))
119 korelacja3=round(cor(dane3),1)
```

```
ggcorrplot(korelacja3,lab = TRUE, title = "23 mm",col=c('red','darkblue','cyan'))
korelacja4=round(cor(dane4),1)
ggcorrplot(korelacja4,lab = TRUE, title = "35 mm",col=c('red','darkblue','cyan'))

### kwadratowa analiza dyskryminacyjna
klasyfikatorQDA = qda(dane[,2:3], grouping = dane[,1], subset=zbior.uczacy)
print(klasyfikatorQDA)
oceny = predict(klasyfikatorQDA, newdata=dane[-zbior.uczacy,2:3])
ct=table(teoretyczne=oceny$class, empiryczne = dane[-zbior.uczacy,1])
ct
diag=(prop.table(ct,1))
sum(diag(prop.table(ct)))
dane$kaliber=as.factor(dane$kaliber)
partimat(factor(dane$kaliber)~dane$czas+dane$odl, data = dane, subset=zbior.uczacy
         ,method = "qda", plot.matrix = FALSE,
         imageplot=TRUE, image.colors=c('cyan', 'deepskyblue','blue','darkblue'),
         col.correct='black', col.wrong='red', main="Klasyfikacja przypadków",
         prec=300, name=c('czas trwania (ms)','odległość czujników (m)'),
         print.err=0)

### regresja logistyczna
dane762=read.csv("C:\\Users\\user\\Desktop\\baza762.csv", header=TRUE, sep=';')
glm.fit=glm(kaliber~odl+czas, data=dane762, subset = zbior.uczacy, family = binomial)
summary(glm.fit)
### test istotności modelu
lrtest(glm.fit)
### walidacja na zbiorze testowym
glm.probs=predict(glm.fit, newdata = dane762[-zbior.uczacy,], type = "response")
Teoretyczne=ifelse(glm.probs>0.5, "1","0")
Empiryczne=dane762$kaliber[-zbior.uczacy]
table(Teoretyczne, Empiryczne)
### procent poprawnej klasyfikacji
mean(Teoretyczne == Empiryczne)
```

```

### KNN
train.X=cbind(dane$odl, dane$czas)[zbior.uczacy,]
test.X=cbind(dane$odl, dane$czas)[-zbior.uczacy,]
train.kaliber=dane$kaliber[zbior.uczacy]
EmpirycznekNN=c(dane$kaliber[-zbior.uczacy])
set.seed(1)
TeoretycznekNN=knn(train.X, test.X, train.kaliber, k=3)
table(TeoretycznekNN,EmpirycznekNN)
mean(TeoretycznekNN==EmpirycznekNN)
```

Załącznik 2. Zestawienie danych

Dane z badań 25 maja 2010 r. [31]			
L.p.	**Kaliber (mm)**	**Czas trwania (ms)**	**Odległość (m)**
1	7,62	0,290	25
2	7,62	0,290	25
3	7,62	0,290	25
4	7,62	0,290	25
5	7,62	0,290	25
6	7,62	0,290	25
7	7,62	0,290	25
8	7,62	0,290	25
9	7,62	0,290	25
10	7,62	0,290	25
11	7,62	0,290	25
12	7,62	0,290	25
13	7,62	0,290	25
14	7,62	0,290	25
15	7,62	0,290	25
16	7,62	0,290	25
17	7,62	0,290	25
18	7,62	0,290	25
19	7,62	0,290	25
20	7,62	0,290	25
21	7,62	0,290	25
22	7,62	0,290	25
23	7,62	0,290	25
24	7,62	0,290	25
25	7,62	0,290	25
26	7,62	0,290	25
27	7,62	0,290	25
28	7,62	0,290	25
29	7,62	0,290	25
30	7,62	0,290	25
31	7,62	0,305	25
32	7,62	0,305	25
33	7,62	0,305	25
34	7,62	0,305	25
35	7,62	0,305	25
36	7,62	0,275	20
37	7,62	0,275	20
38	7,62	0,275	20
39	7,62	0,275	20
40	7,62	0,275	20
41	7,62	0,275	20
42	7,62	0,275	20
43	7,62	0,275	20
44	7,62	0,275	20
45	7,62	0,275	20
46	7,62	0,275	20
47	7,62	0,275	20
48	7,62	0,275	20
49	7,62	0,290	20
50	7,62	0,290	20
51	7,62	0,290	20
52	7,62	0,290	20
53	7,62	0,290	20
54	7,62	0,290	20
55	7,62	0,290	20
56	7,62	0,290	20
57	7,62	0,290	20
58	7,62	0,290	20
59	7,62	0,290	20
60	7,62	0,290	20
61	7,62	0,290	20
62	7,62	0,290	20
63	7,62	0,290	20
64	7,62	0,290	20
65	7,62	0,290	20
66	7,62	0,290	20
67	7,62	0,290	20
68	7,62	0,290	20
69	7,62	0,275	15
70	7,62	0,275	15
71	7,62	0,275	15
72	7,62	0,275	15
73	7,62	0,275	15
74	7,62	0,275	15
75	7,62	0,275	15
76	7,62	0,275	15
77	7,62	0,275	15
78	7,62	0,275	15

79	7,62	0,275	15
80	7,62	0,275	15
81	7,62	0,275	15
82	7,62	0,275	15
83	7,62	0,275	15
84	7,62	0,275	15
85	7,62	0,275	15
86	7,62	0,275	15
87	7,62	0,275	15
88	7,62	0,275	15
89	7,62	0,275	15
90	7,62	0,275	15
91	7,62	0,275	15
92	7,62	0,275	15
93	7,62	0,275	15
94	7,62	0,275	15
95	7,62	0,275	15
96	7,62	0,275	15
97	7,62	0,275	15
98	7,62	0,275	15
99	7,62	0,275	15
100	7,62	0,275	15
101	7,62	0,275	15
102	7,62	0,275	15
103	7,62	0,244	10
104	7,62	0,259	10
105	7,62	0,259	10
106	7,62	0,259	10
107	7,62	0,259	10
108	7,62	0,259	10
109	7,62	0,259	10
110	7,62	0,259	10
111	7,62	0,259	10
112	7,62	0,259	10
113	7,62	0,259	10
114	7,62	0,259	10
115	7,62	0,259	10
116	7,62	0,259	10
117	7,62	0,259	10
118	7,62	0,259	10
119	7,62	0,259	10
120	7,62	0,259	10
121	7,62	0,259	10
122	7,62	0,214	7,5
123	7,62	0,214	7,5
124	7,62	0,214	7,5
125	7,62	0,214	7,5
126	7,62	0,214	7,5
127	7,62	0,214	7,5
128	7,62	0,214	7,5
129	7,62	0,214	7,5
130	7,62	0,214	7,5
131	7,62	0,214	7,5
132	7,62	0,214	7,5
133	7,62	0,214	7,5
134	7,62	0,214	7,5
135	7,62	0,214	7,5
136	7,62	0,214	7,5
137	7,62	0,214	7,5
138	7,62	0,214	7,5
139	7,62	0,214	7,5
140	7,62	0,214	7,5
141	7,62	0,214	7,5
142	7,62	0,214	7,5
143	7,62	0,214	7,5
144	7,62	0,214	7,5
145	7,62	0,214	7,5
146	7,62	0,214	7,5
147	7,62	0,214	7,5
148	7,62	0,229	7,5
149	7,62	0,229	7,5
150	7,62	0,198	2,5
151	7,62	0,198	2,5
152	7,62	0,198	2,5
153	7,62	0,198	2,5
154	7,62	0,198	2,5
155	7,62	0,198	2,5
156	7,62	0,198	2,5
157	7,62	0,198	2,5
158	7,62	0,198	2,5
159	7,62	0,198	2,5
160	7,62	0,198	2,5
161	7,62	0,198	2,5
162	7,62	0,198	2,5
163	7,62	0,198	2,5
164	7,62	0,198	2,5
165	7,62	0,198	2,5
166	7,62	0,198	2,5
167	7,62	0,198	2,5
168	7,62	0,198	2,5

169	7,62	0,198	2,5
170	7,62	0,198	2,5
171	7,62	0,198	2,5
172	7,62	0,198	2,5
173	7,62	0,198	2,5
174	7,62	0,198	2,5
175	7,62	0,198	2,5

176	7,62	0,198	2,5
177	7,62	0,198	2,5
178	7,62	0,198	2,5
179	7,62	0,198	2,5
180	7,62	0,198	2,5
181	7,62	0,198	2,5

Dane z badań 16 grudnia 2006 r. [32]			
L.p.	**Kaliber (mm)**	**Czas trwania (ms)**	**Odległość (m)**
1	7,62	0,164	1,25
2	7,62	0,160	1,25
3	7,62	0,150	0,85
4	7,62	0,150	0,85
5	7,62	0,146	0,85
6	7,62	0,146	0,85
7	7,62	0,144	0,85
8	7,62	0,144	0,85
9	7,62	0,143	0,85
10	7,62	0,120	0,35
11	5,56	0,120	1,22
12	7,62	0,118	0,35
13	7,62	0,112	0,25
14	7,62	0,110	0,25
15	5,56	0,110	1,03
16	5,56	0,110	1,03
17	5,56	0,108	1,03
18	5,56	0,108	1,03
19	5,56	0,100	0,7
20	5,56	0,098	0,7
21	5,56	0,092	0,7
22	5,56	0,090	0,7
23	5,56	0,090	0,7

Dane z badań 14 grudnia 2018 r. [33]			
L.p.	**Kaliber (mm)**	**Czas trwania (ms)**	**Odległość (m)**
1	23	0,481	3,276
2	23	0,536	3,276
3	23	0,484	3,191
4	23	0,496	3,191
5	23	0,525	3,140
6	23	0,472	3,140
7	23	0,495	3,108
8	23	0,536	3,108
9	23	0,483	3,089
10	23	0,574	3,089
11	23	0,487	3,089
12	23	0,522	3,089
13	23	0,520	3,089
14	23	0,520	3,089
15	23	0,521	2,997
16	23	0,533	2,997
17	23	0,503	1,965
18	23	0,431	1,965
19	23	0,439	1,726
20	23	0,458	1,726
21	23	0,415	1,712
22	23	0,460	1,712
23	23	0,414	1,530
24	23	0,421	1,530
25	23	0,425	1,530
26	23	0,425	1,530
27	23	0,442	1,530
28	23	0,428	1,530
29	23	0,440	1,365
30	23	0,456	1,365
31	23	0,454	1,334
32	23	0,442	1,334

Dane z badań 9 stycznia 2018 r. [34]			
L.p.	**Kaliber (mm)**	**Czas trwania (ms)**	**Odległość (m)**
1	35	0,831	11
2	35	0,866	11
3	35	0,868	11
4	35	0,888	11
5	35	0,841	11
6	35	0,873	11
7	35	0,873	11
8	35	0,885	11
9	35	0,808	11
10	35	0,827	11
11	35	0,821	11
12	35	0,834	11
13	35	0,812	11
14	35	0,837	11
15	35	0,849	11
16	35	0,844	11
17	35	0,725	6,284
18	35	0,753	6,284
19	35	0,784	6,284
20	35	0,676	6,284
21	35	0,702	6,284
22	35	0,738	6,284
23	35	0,713	6,284
24	35	0,746	6,284
25	35	0,730	6,284
26	35	0,741	6,284
27	35	0,698	6,284
28	35	0,726	6,284
29	35	0,714	6,284
30	35	0,726	6,284
31	35	0,689	6,284
32	35	0,681	6,284
33	35	0,684	6,284
34	35	0,657	6,284
35	35	0,611	3,490
36	35	0,621	3,490
37	35	0,632	3,490
38	35	0,645	3,490
39	35	0,594	3,490
40	35	0,608	3,490
41	35	0,620	3,490
42	35	0,643	3,490

(strona celowo zostawiona pusta)

Załącznik nr 6
do „Szczegółowych zasad oraz harmonogramu
wykonywania prac dyplomowych na
Wydziale Mechatroniki i Lotnictwa WAT”

Warszawa, ... 201… r.

Wojskowa Akademia Techniczna
Wydział Mechatroniki i Lotnictwa
Imię i nazwisko studenta: sierż. pchor. inż. Dawid Brejecki
Numer albumu: 62405

OŚWIADCZENIE

Świadomy(a) odpowiedzialności karnej z tytułu naruszenia przepisów ustawy o prawie autorskim i prawach pokrewnych (Dz. U. nr 80 poz. 904 z 2000 r. ze zmianami) oraz konsekwencji dyscyplinarnych określonych w ustawie z dnia 27 lipca 2005 r. prawo o szkolnictwie wyższym (Dz.U. Nr 164 poz. 1365, z późn. zm.)[1)], a także odpowiedzialności cywilnoprawnej oświadczam, że przedkładana praca dyplomowa pt.
„Metody statystyczne w klasyfikacji pocisków na podstawie danych pochodzących ze strzelań”
została napisana przeze mnie samodzielnie i nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadaniem dyplomu uczelni lub tytułów zawodowych. Jednocześnie oświadczam, że wyżej wymieniona praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy o prawie autorskim i prawach pokrewnych innych osób oraz dóbr osobistych chronionych prawem cywilnym. Wszystkie informacje umieszczone w pracy, uzyskane ze źródeł pisanych i elektronicznych oraz inne informacje, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami.

.....
/data, podpis studenta/

[1)] Ustawa z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym:
Art. 214 ust. 4. „W razie podejrzenia popełnienia przez studenta czynu polegającego na przypisaniu sobie autorstwa istotnego fragmentu lub innych elementów cudzego utworu rektor niezwłocznie poleca przeprowadzenie postępowania wyjaśniającego.”;
Art. 214 ust. 6. „Jeżeli w wyniku postępowania wyjaśniającego zebrany materiał potwierdza popełnienie czynu, o którym mowa w ust. 4, rektor wstrzymuje postępowanie o nadanie tytułu zawodowego do czasu wydania orzeczenia przez komisję dyscyplinarną oraz składa zawiadomienie o popełnieniu przestępstwa.”.

Wyrażam zgodę na udostępnienie mojej pracy dyplomowej przez Archiwum WAT w czytelni i w ramach wypożyczeń międzybibliotecznych.

.....
/data, podpis studenta/