Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Wydział Elektroniki i Technik Informacyjnych

Projekt nr. 36: Diagnostyka raka piersi w badaniach mammograficznych za pomocą sieci SOM

(katalog: Mammographic Mass_MLR)

Sieci neuronowe w zastosowaniach biomedycznych
na kierunku Inżynieria biomedyczna

Kacper Kilianek (305375), Adam Piszczek (303803)
Zespół nr. 22

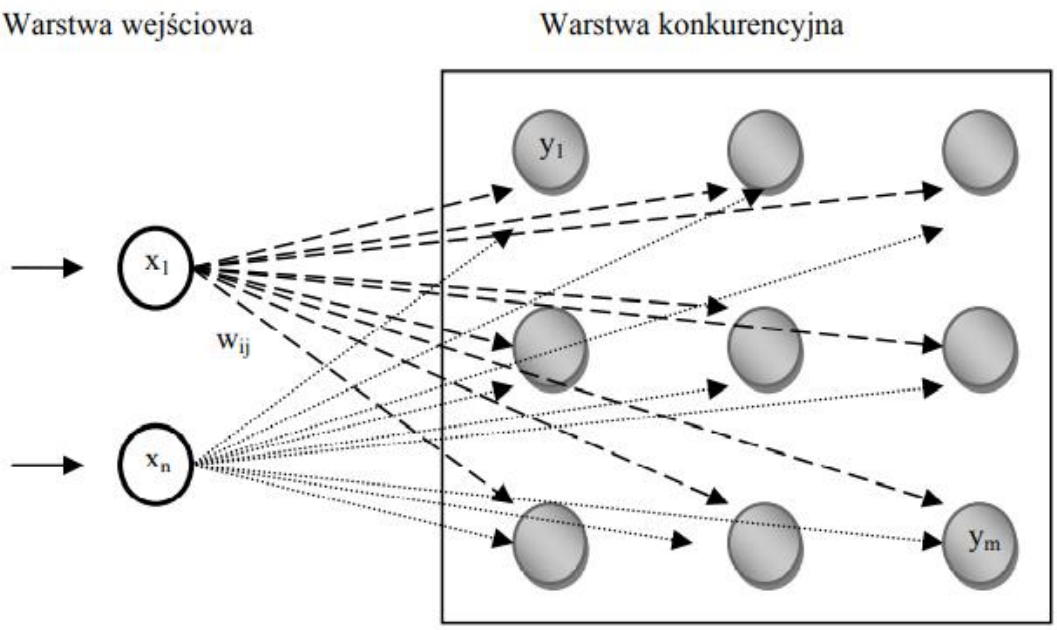
Prowadzący projekt
dr inż. Paweł Mazurek

Warszawa 14.04.2022 r.

Spis treści

Wprowadzenie

W niniejszym projekcie korzystać będziemy z sieci neuronowych typu SOM (ang. Self-organizing Maps). Jak wynika to z angielskiej nazwy, są to sieci samoorganizujące się. Z tego powodu będziemy mieć do czynienia z uczeniem bez nadzoru, całkiem inaczej niż w przypadku alternatywnej sieci MLP. Co to oznacza? W przypadku „sieci nienadzorowanych" podczas treningu nie są przedstawiane żadne wzorce wyjścia. „Zadaniem sieci jest dopiero stworzenie takich wzorców podczas etapu uczenia się. Dane treningowe są samodzielnie klasyfikowane przez sieć jedynie na podstawie ich wzajemnej korelacji. Trenowanie odbywa się więc jako tzw. proces samouczenia. (…) Sieci SOM zbudowane są z dwóch warstw: warstwy wejściowej oraz warstwy wyjściowej, nazywanej też warstwą konkurencyjną…" [1]. Informacją wartą zauważenia jest to, iż w przeciwieństwie do innych rodzajów SN, sieci SOM nie zawierają warstwy ukrytej. „Neurony pierwszej warstwy nie dokonują żadnych przekształceń danych, a mają za zadanie jedynie rozesłać wszystkie wartości wprowadzone na wejścia sieci do warstwy konkurencyjnej" [1]. Wadą sieci SOM jest to, iż przy modyfikacji istniejących danych lub dodaniu do nich porcji nowych, nauka sieci musi zostać powtórzona od początku.



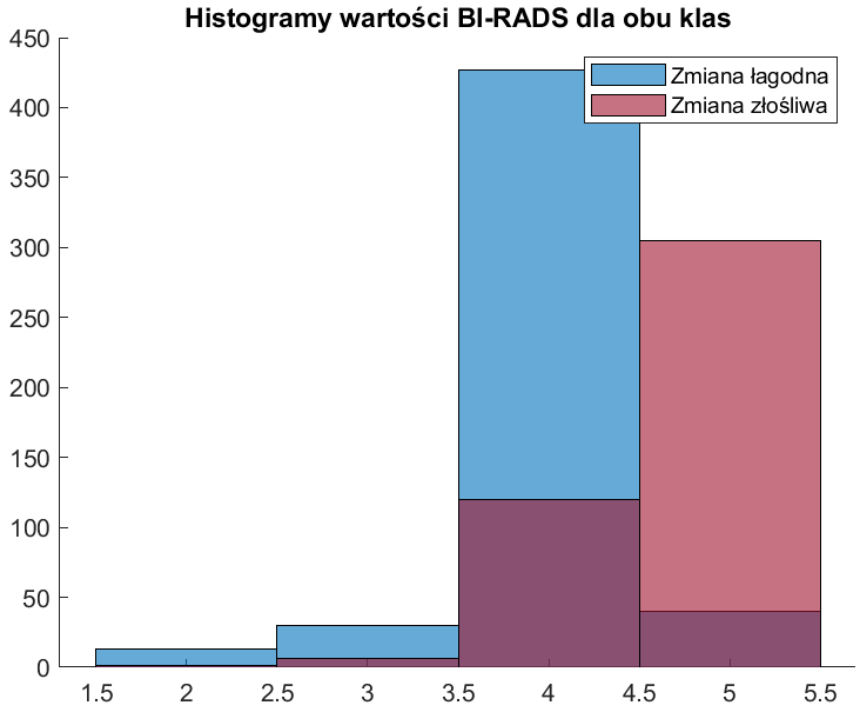
$x_1, \ldots, x_n$ – neurony warstwy wejściowej,
$y_1, \ldots, y_m$ – neurony warstwy wyjściowej,
w_{ij} – j-ta wartość wektora wag i-tego neuronu warstwy wyjściowej
($i = 1, \ldots, m$; $j = 1, \ldots, n$),
n – liczba wejść,
m – liczba wyjść.

Rysunek 1- Architektura sieci SOM [1]

Wstępna analiza danych

W naszym projekcie korzystamy ze zbioru danych o nazwie Mammographic Mass_MLR. Są to dane dotyczące wykonanych badań mammograficznych na przełomie lat 2003 i 2006. Dane te zawierają informacje:

1. Ocenę BI-RADS – jest to system standaryzujący opis mammograficzny, wyniki w jego skali są od 1 do 5; im wyższa liczba, tym większe ryzyko złośliwości zmiany



Rysunek 2 - Histogram wartości BI-RADS dla obu klas [6]

Warto w tym miejscu dodać, dlaczego nie możemy dojrzeć na naszym histogramie wartości 1 – otóż w naszym zbiorze danych minimalną wartością jest 2, dlatego też wartości mniejsze od niej nie są uwzględnione. Na potwierdzenie umieszczamy otrzymane zliczenia z programu MATLAB:

Label	Count	Percent
2	7	0.858895705521472
3	24	2.94478527607362
4	468	57.4233128834356
5	316	38.7730061349693

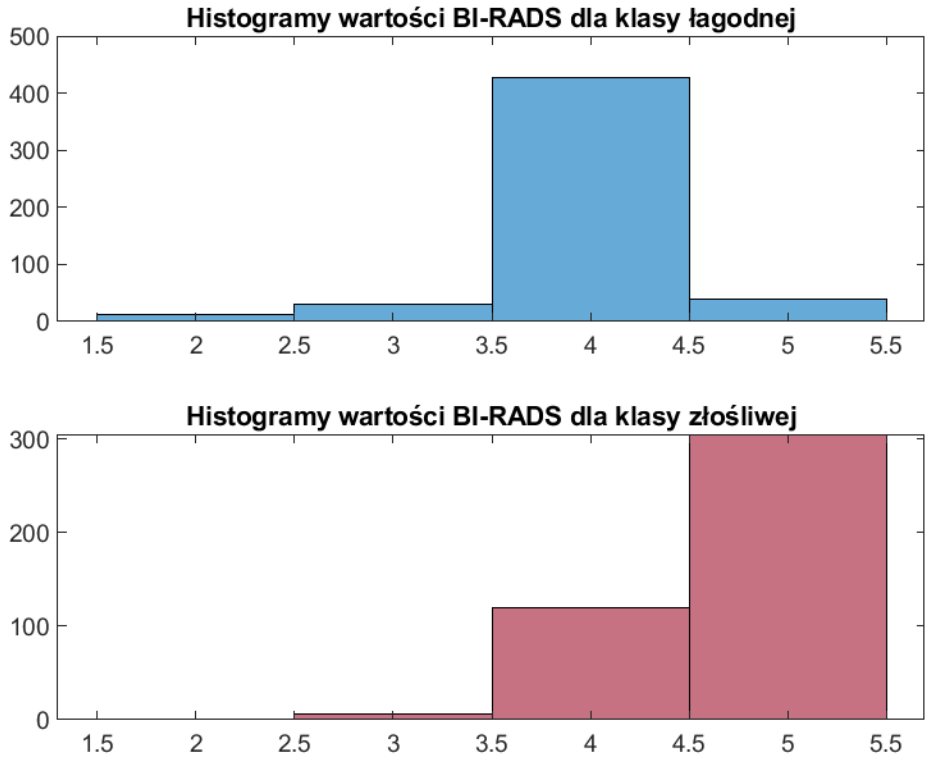
W tym miejscu można również zauważyć, iż największy udział przypada wartości równej 4, która określana jest jako zmiany podejrzane:

Tabela 1 - Kategoria opisu zmian według BI-RADS [2]

4	zmiana podejrzana	ryzyko złośliwości od 2% do 95%, konieczna weryfikacja zmiany, grupę podzielono na 3 podgrupy(*): • 4a: zmiana podejrzana, ale o małym stopniu prawdopodobieństwa złośliwości • 4b: zmiana podejrzana, o pośrednim stopniu prawdopodobieństwa złośliwości • 4c: zmiana podejrzana, o wysokim stopniu prawdopodobieństwa złośliwości, jednak bez klasycznych cech złośliwości

Dodatkowe parametry opisujące ten atrybut to średnia i odchylenie standardowe. Wynoszą one odpowiednio 4.298301486199575 oraz 0.612634667039133. Na podstawie wartości średniej możemy zauważyć większą dominację zdarzeń równych 4 jak i zarówno dość mały rozrzut danych.

Następnie przyjrzyjmy się wykresom dla obu klas z osobna:



Rysunek 3 - Porównanie histogramów wartości BI-RADS [6]

Jak widzimy powyżej, znaczna większość wyników dla klasy złośliwej otrzymała wynik 5, świadczący o wysokim prawdopodobieństwie złośliwości:

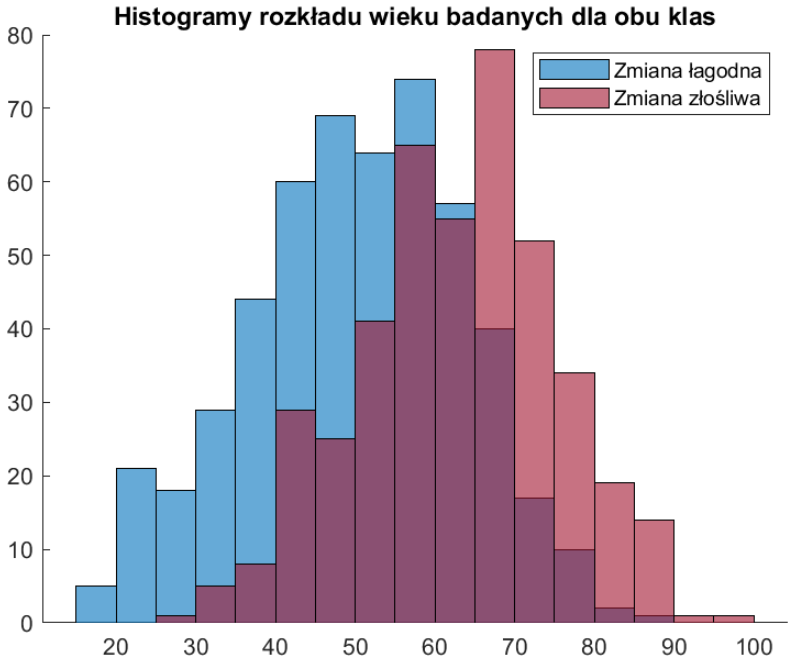
Tabela 2 - Kategoria opisu zmian według BI-RADS [2]

5	zmiana o wysokim prawdopodobieństwie złośliwości	ryzyko złośliwości >95%, konieczna weryfikacja zmiany i dalsze leczenie

Jeśli chodzi o parametry dla naszych dwóch klas są one następujące:

- Dla klasy łagodnej – średnia wynosi 3.968627450980392, odchylenie standardowe 0.488568815678619
- Dla klasy złośliwej – średnia 4.6875, odchylenie standardowe 0.507055784352645

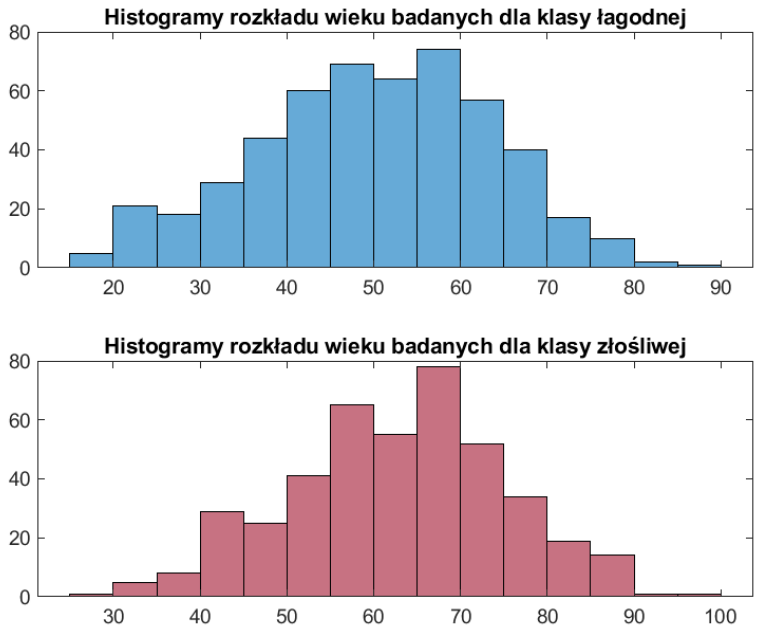
2. Wiek pacjentki



Rysunek 4 - Histogram rozkładu wieku badanych [6]

Średnia wieku pacjentek wynosi ok. 55 lat a odchylenie standardowe ich wieku to 14.487758155334294.

Kolejnym etapem analizy jest pochylenie się nad histogramami dla obu klas z osobna:



Rysunek 5 - Porównanie histogramów wieku pacjentek [6]

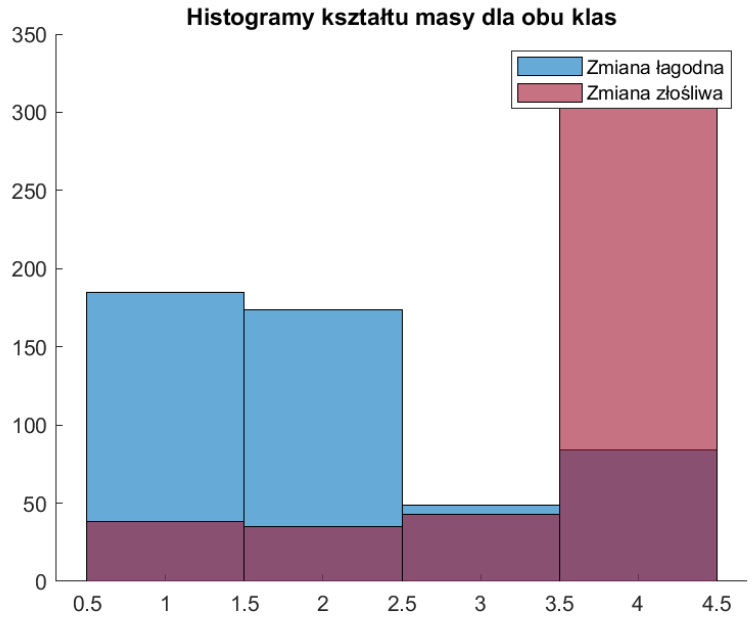
Jak można powiedzieć, wykresy te są trochę przesunięte w stosunku do siebie – więcej zmian złośliwych dotyczy kobiet starszych niż jest to w przypadku zmian łagodnych. Jednak należy podkreślić, że sam wiek nie jest zmienną decyzyjną, a jest jednym z atrybutów, które analizowane są przez sieć SOM w celu wydania klasyfikacji.

W przypadku tej cechy, średnia i odchylenie standardowe wynoszą:

- Dla klasy łagodnej - 49.645792563600786 oraz 13.628753374292089
- Dla klasy złośliwej - 62.273364485981311 oraz 12.339177753127601

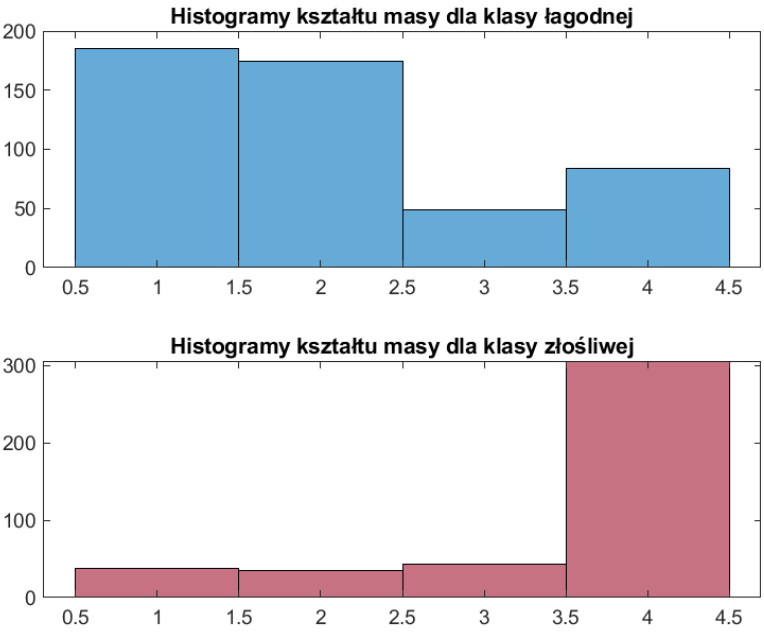
Co potwierdza wyżej sformułowany wniosek.

3. Kształt zmiany – oznaczany w zbiorze danych liczbowo od 1 do 4, odpowiadający oznaczeniom: okrągły, owalny, zrazikowy, nieregularny



Rysunek 6 - Histogram kształtu masy [6]

Dla wszystkich wyników badań zanotowano średnią równą 2.710065645514223 oraz odchylenie standardowe równe 1.244557657758133. Średnią dla tej cechy mocno zawyżają wyniki dla klasy złośliwej, co jest widoczne poniżej:

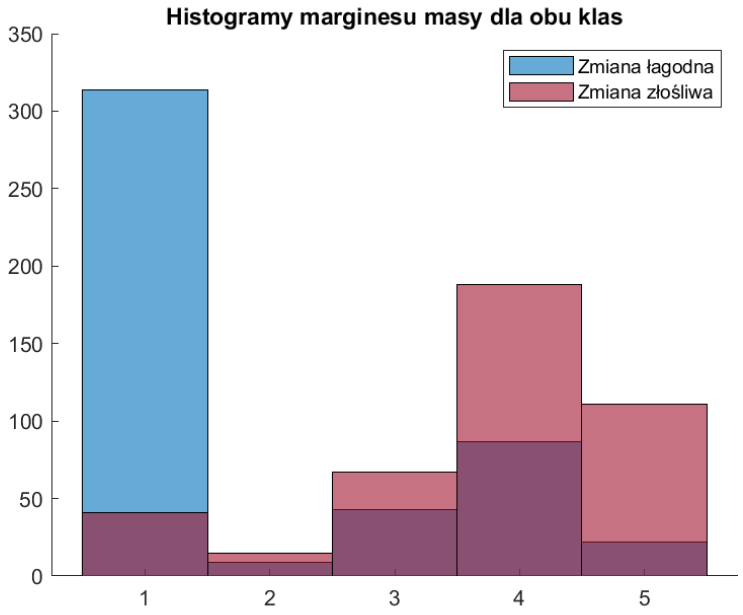


Rysunek 7 - Porównanie histogramów kształtu masy [6]

Ogromna większość zmian złośliwych cechuje się nieregularnym kształtem, pozostałe wyniki są prawie pomijalne. W przypadku zmian łagodnych większość z nich cechuje się kształtem okrągłym bądź owalnym.

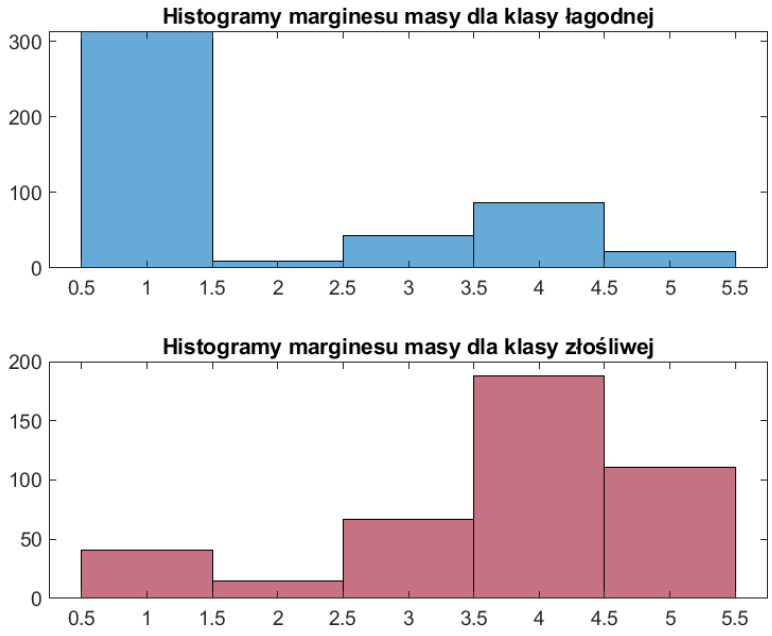
Parametry dla obu klas wynoszą 2.065040650406504 (średnia) oraz 1.075479999110511 (odch. standardowe) dla klasy łagodnej oraz 3.462085308056872 (średnia) oraz 0.978259354990722 (odch. standardowe) dla klasy złośliwej. Wyniki dla klasy złośliwej charakteryzują się mniejszą zmiennością.

4. Margines zmiany - oznaczany w zbiorze danych liczbowo od 1 do 5, odpowiadający oznaczeniom: ograniczony, microlobulated, zasłonięty, niewyraźny, spikularny



Rysunek 8 - Histogram marginesu masy dla obu klas [6]

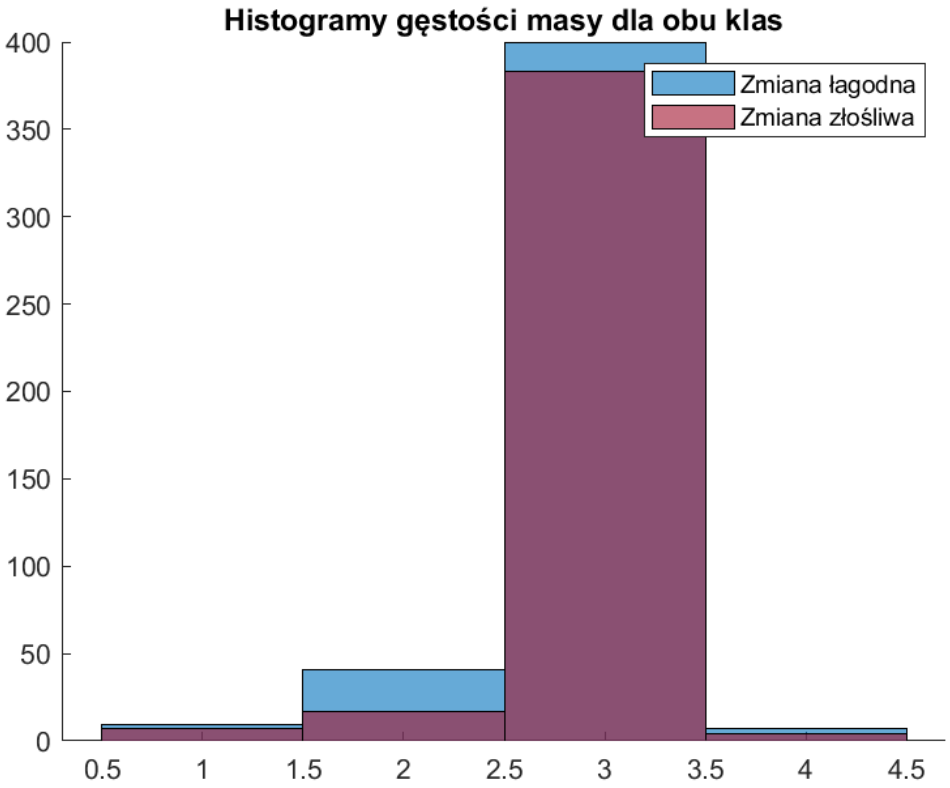
Sytuacja zawyżania wyniku przez klasę złośliwą również się powtarza w przypadku marginesu zmiany. Dla ogółu średnia wynosi 2.784838350055741, a odchylenie standardowe 1.570328176560257. Jeśli chodzi o nasze klasy, parametry te wynoszą odpowiednio 1.934736842105263 oraz 1.378693231120878 dla klasy łagodnej oraz 3.741706161137441 i 1.172574538716222 dla klasy złośliwej.



Rysunek 9 - Porównanie histogramów marginesu masy [6]

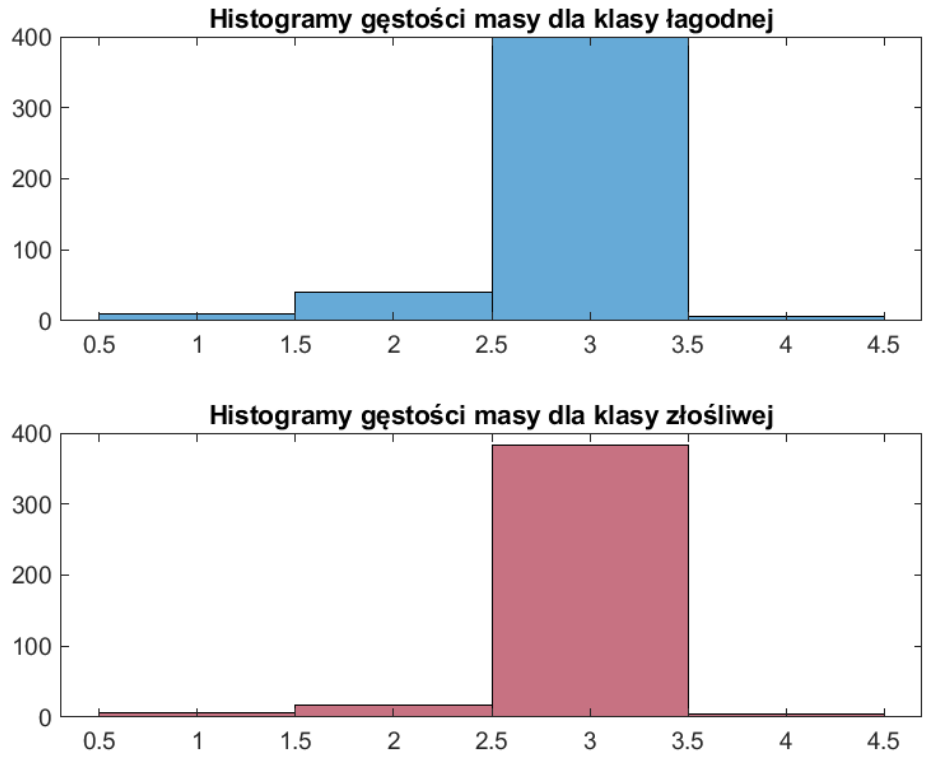
Zmiany złośliwe najczęściej cechuje margines niewyraźny lub spikularny, zaś zmiany łagodne margines ograniczony. Im w zbiorze danych wystąpi więcej tego typu korelacji, pomoże to sieci na utworzenie dokładniejszego schematu postępowań z nowymi (nieznanymi sieci) danymi wejściowymi.

5. Gęstość zmiany - oznaczana w zbiorze danych liczbowo od 1 do 4, odpowiadający oznaczeniom: wysoka, izo-, niska, zawierająca tłuszcz



Rysunek 10 - Histogram gęstości masy [6]

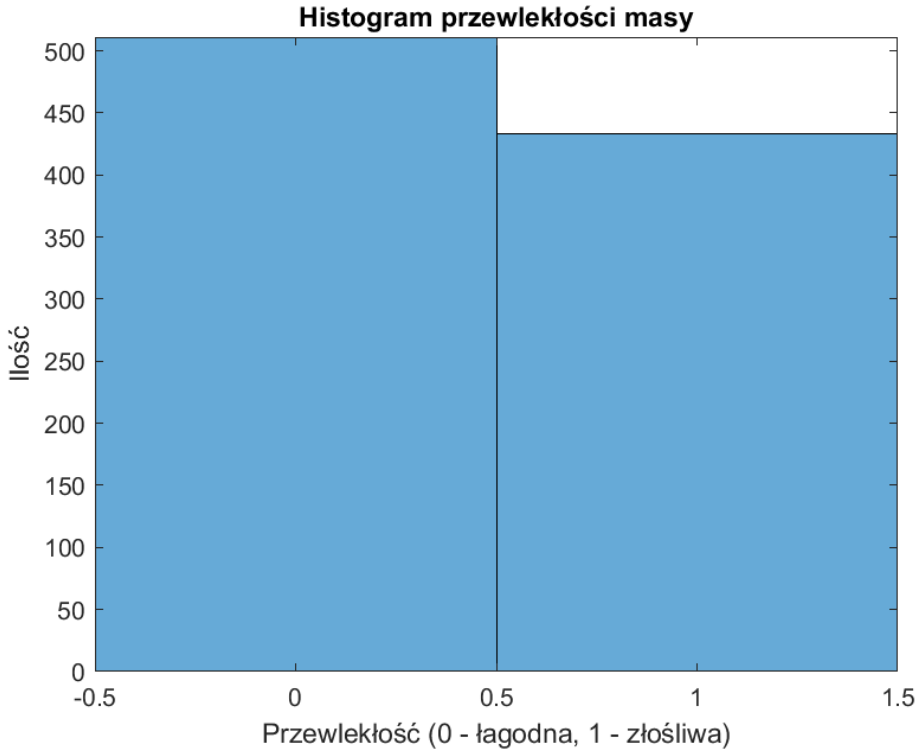
Na powyższym wykresie widoczna jest znaczna dominacja niskiej gęstości masy pośród wszystkich zanotowanych wyników. Zatem jest to jedna z cech, która daję nam niski stopień rozróżnialności pomiędzy klasami. Jest to potwierdzone przez nasze parametry informacyjne – średnią oraz odchylenie standardowe. Dla wszystkich wyników wyniosły one 2.908986175115207 oraz 0.380932365304760, dla klasy łagodnej 2.886214442013129 i 0.413805485982928, a dla klasy złośliwej 2.934306569343066 oraz 0.339390692817839. Jak możemy łatwo zauważyć, wszystkie wartości tychże parametrów w przypadku tej cechy są bardzo zbliżone do siebie.



Rysunek 11 - Porównanie histogramów gęstości masy [6]

Niestety w przypadku gęstości masy, nie moglibyśmy tylko na jej podstawie stwierdzić czy dana zmiana z większym prawdopodobieństwem jest złośliwa lub łagodna, gdyż obie klasy charakteryzuje najczęściej niską gęstość masy. Wynika to wprost z niemalże identycznego rozkładu wartości tej cechy co powoduje, że jest ona mało pomocna w kontekście samej diagnozy i nie można się na niej bezpośrednio opierać.

6. Rodzaj zmiany – łagodna lub złośliwa



Rysunek 12 - Histogram przewlekłości masy [6]

Jeśli chodzi o zdiagnozowane zmiany, przeważająca większość z nich jest łagodna. Warto mieć to na uwadze podczas przeprowadzania wstępnego przetwarzania danych, gdyż nie chcemy dopuścić do sytuacji, gdzie trenowana sieć SOM zacznie faworyzować jedną ze stron, która liczniej występuje w zbiorze danych.

Wybrana metoda wstępnego przetwarzania danych (normalizacja, oraz obróbka niepotrzebnych indeksów cech)

W przypadku naszego zestawu danych składa się on z 961 liczby badań, które zostały ściślej przedstawione w poprzednim rozdziale. Jednak podczas analizy danych okazało się, że część z nich jest niekompletna (w miejscu pomiaru/danej widnieje znak '?' lub też jest błędnie wpisana wartość spoza przyjętego zakresu). Aby nasza sieć samoorganizująca przeprowadziła dobry proces uczenia, postanowiliśmy usunąć wszelkie błędne dane. Głównym warunkiem odrzucenia wierszy była obecność wartości wykraczających poza zakres jak np. wartość 55 dla cechy BI-RADS, która przecież jest z przedziału 1-5. Taka wstępna obróbka pozwala na odrzucenie tzw. z ang. outliers (danych odstających), co ostatecznie wpłynęło na redukcje 17 wierszy danych, pozostawiając już 944 wiersze o lepszych możliwościach klasyfikacyjnych. Co do nieznanych wartości są one pozostawione ze względu na wybrany sposób liczenia odległości oparty na entropii. Zapewnia on spójne podejście do obsługi zarówno nominalnych, jak i liczbowych atrybutów, a dodatkowo umożliwia matematycznie dobrze ugruntowane podejście do obsługi brakujących wartości [7].

W naszym przypadku jest to bardzo istotna zależność, ponieważ niekompletnych wektorów jest aż 146, co przekłada się na około 15% całego dostępnego zbioru danych.

Dodatkowo w celu polepszenia i usprawnienia procesu uczenia sieci, jak lepszego zrozumienia problemu związanego z pracą na pięciowymiarowym zestawie danych, postanowiliśmy znormalizować nasze dane do skali 0-1. W przypadku naszej sytuacji ma to o tyle znaczenie, że jedna z cech określająca wiek jest z zakresu 18-96, gdzie pozostałe cechy plasują się raczej w zakresie wartości 1-4, bądź 1-5. Gdyby dane nie zostały poprzedzone normalizacją mogłoby dojść do sytuacji, gdzie przy liczeniu odległości pomiędzy nimi największy wpływ na klasyfikację miałaby cecha związana z wiekiem, ponieważ jej wartości są o rząd wielkości większe od pozostałych atrybutów. W dystansach euklidesowych, które wykorzystamy w implementacji naszej sieci, dominuje próbka o największej wartości, zatem nie należy polegać na odległości euklidesowej, jeżeli mamy do czynienia z danymi nieznormalizowanymi [4]. Normalizacje do przedziału [0,1] przeprowadziliśmy na podstawie poniższego wzoru:

$$B = \frac{(A - \min(A))}{(\max(A) - \min(A))}(D - C) + C \tag{1}$$

gdzie:

B - jest znormalizowanym wektorem cechy

A – wektor cechy przed normalizacją

D – dolny zakres pożądanego przedziału danych, równy 0

C – górny zakres pożądanego przedziału danych, równy 1

Co do samego podziału zebranych atrybutów na dane uczące i testowe postanowiliśmy rozdzielić je na dwa zestawy, przy czym zachowując w miarę równość pomiędzy ilościami złośliwego i łagodnego znamienia nowotworu w obu zestawach. Tak przeprowadzona analiza danych pozwoli sieci SOM na stworzenie własnych wzorców, które nie będą faworyzować żadnej ze stron i zostanie przeprowadzona obiektywniejsza klasyfikacja nowotworu.

Koncepcja realizacji sieci neuronowej

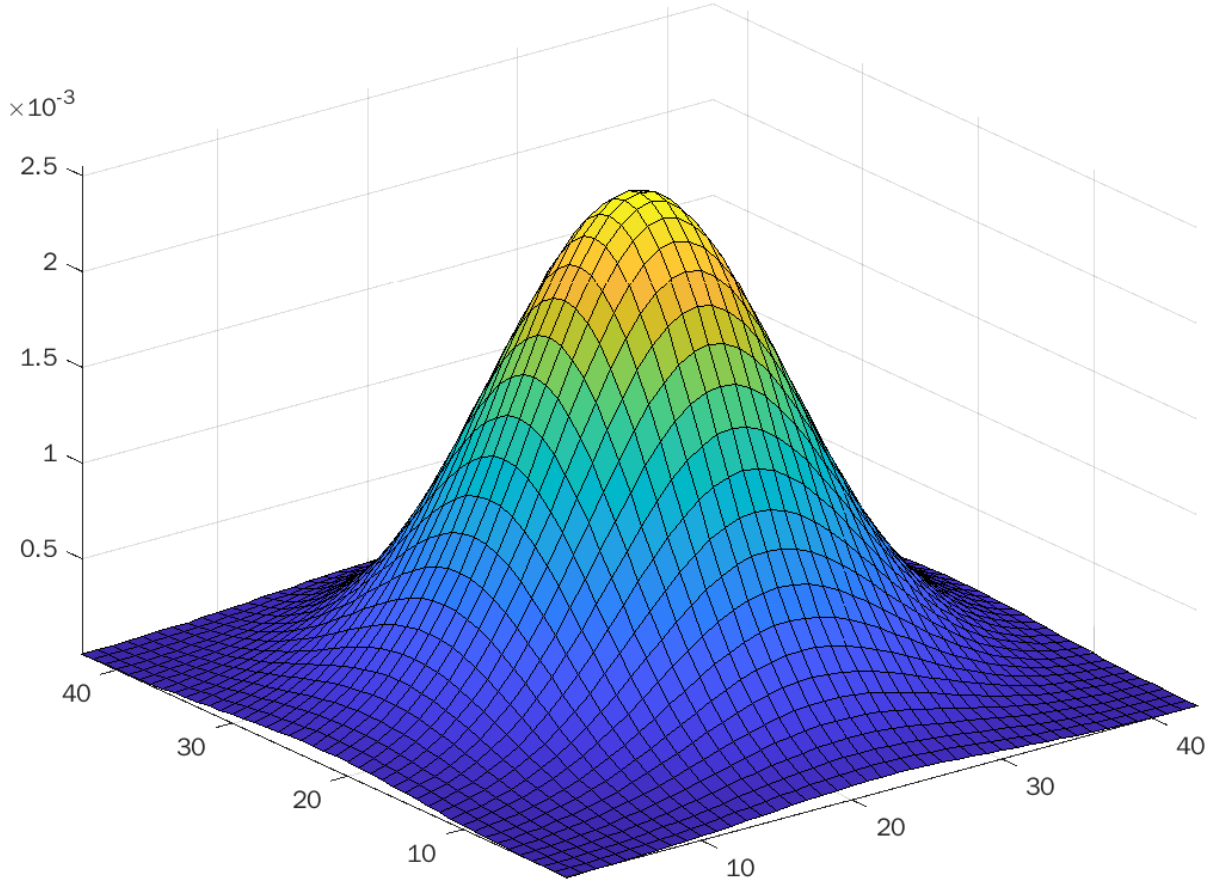
Do realizacji naszej sieci SOM postanowiliśmy utworzyć sieć składającą się z 36 neuronów, tworzących siatkę 6x6. Poniżej zamieściliśmy schemat takiej sieci, które ilustruje nam zamysł związany z przetwarzaniem danych i relacjami jakimi powiązane są dane neurony ze sobą i z wejściowymi wektorami danych. Co do samej funkcji aktywacji (sąsiedztwa) stwierdziliśmy, że lepszym rozwiązaniem będzie zaprzęgnięcie funkcji gaussowskiej w zależności, od której powiązane neurony w okolicy zaktywizowanego neuronu będą również aktywowane z odpowiednim pobudzeniem (zależnym od odległości). Sam sposób wyznaczenia tego typu funkcji oraz jej trójwymiarowa wizualizacja zostały przedstawione poniżej:

$$h(m, m^*) = \exp\left(-\frac{\|\boldsymbol{r}_m - \boldsymbol{r}_{m^*}\|_2^2}{2[\sigma(t)]^2}\right) \tag{2}$$

gdzie:

$\boldsymbol{r}_m i\, \boldsymbol{r}_{m^*}$ - oznaczają wektory określające położenie neuronów m oraz m^*

$\sigma(t)$ - oznacza odchylenie standardowe rozkładu określające jego zasięg; z czasem parametr zmniejsza swoją wartość.



Rysunek 13 – wizualizacja funkcji Gaussa [6]

Przechodząc do kwestii samego algorytmu uczenia naszej sieci WTM, należy podkreślić, że podczas samoorganizowania się sieci, modyfikowane są wagi w sąsiedztwie zapalanego neuronu, a nie tak jak ma to miejsce w WTA, gdzie modyfikowany są tylko wagi zwycięskiego neuronu. Sama intensywność zmian określona jest przez tzw. funkcję sąsiedztwa, która w naszym przypadku przyjmuje tak jak wcześniej wspomniano funkcję Gaussa. Kolejnym istotnym elementem jest współczynnik uczenia, który w naszym przypadku będzie malejący w czasie, gdyż dążymy do tego aby algorytm szybko trafił w optymalne miejsce, żeby potem mniejszymi krokami podążał w kierunku optymalnych wartości wag. Omawianą zależność możemy przedstawić według następującego równania [2]:

$$\boldsymbol{w}_m(t+1) = \boldsymbol{w}_m(t) + \eta(t) \cdot h(m, m^*) \cdot [\boldsymbol{x}^k - \boldsymbol{w}_m(t)] \qquad (3)$$

gdzie:

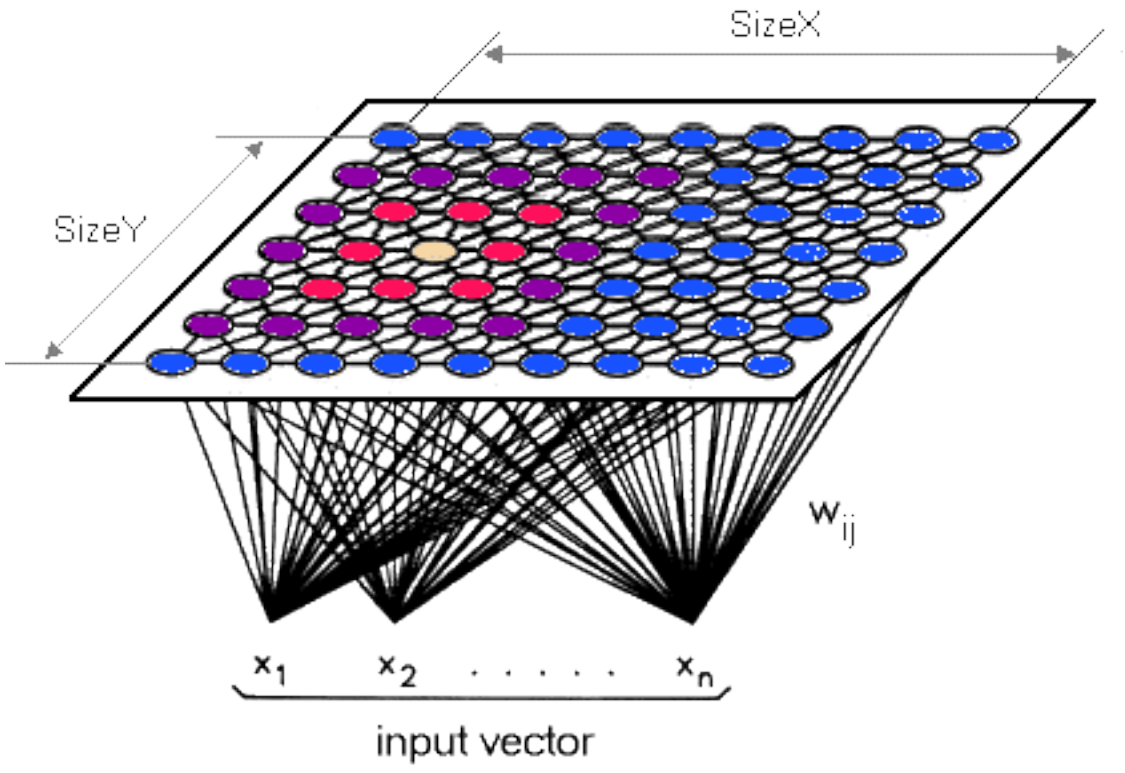
m^* – oznacza indeks neuronu zwycięskiego

$\eta(t)$ – współczynnik uczenia sieci (systematycznie zmniejszający się w czasie t)

Jak można zauważyć zarówno współczynnik uczenia jak i wektor wag będzie się zmieniał w kolejnych krokach. Dodatkowo w przypadku procesu uczenia wektor wag będzie inicjalizowany losowo, jak również

przypadkowo będzie wybierany dany wektor cech na wejście algorytmu sieci (w każdym kolejnym kroku). Zatem w ogólności sam proces uczenia będzie związany ze zmianą struktury sieci, która to w kolejnych iteracjach będzie kształtowała się tak, aby uwydatnić dane cechy i w efekcie ułatwić sobie wydanie werdyktu. Sam proces z definicji przeprowadzany jest bez jakiegokolwiek nadzoru i jedynym testem, czy sieć dobrze sobie poradziła, będzie końcowe porównanie jej efektów z dostępnymi wynikami danych. Dobrana koncepcja pozwoli na zoptymalizowanie nauki sieci, jak i poprawę jej efektów klasyfikacji, gdyż taka ilość neuronów umożliwi obiektywne i dokładne reagowanie na podobne wzorce wejściowe [3].

Jeśli chodzi o samą koncepcję realizacji sposobu klasyfikowania odpowiedzi wyjściowej, postanowiliśmy skorzystać z podpowiedzi autorów artykułu „The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process". Przedstawili oni swoje podejście, w którym decyzja związana z ustaleniem grupy danego wektora cech jest realizowana na podstawie podobieństwa wartości zbiorów. Tworzona jest zmienna decyzyjna q, która w zależności od wartości podobieństwa względem całego zbioru danych i jego N regionów, wybiera k-tą ilość regionów i liczony jest wtedy ułamek f_m złośliwych regionów. Ustawiany jest próg decyzyjny C_f (z zakresu od 0 do 1), a przypadek dla nowotworu złośliwego egzekwowany jest w przypadku $f_m \geq C_f$ (inaczej klasyfikowana jest zmiana łagodna) [7].



Rysunek 14 – schemat sieci samoorganizujące się sieci SOM [3]

Podsumowanie

Rozważana samoorganizująca się sieć SOM jest idealna do klasyfikacji obiektów, które mogą być oceniane na podstawie wzorców, niewidocznych na pierwszy rzut oka. Dzięki znalezieniu zależności pomiędzy grupami cech, sieć ta jest w stanie tworzyć wzorce, na podstawie których później wydaje obiektywne klasyfikacje danych wejściowych. Dzięki temu są one szeroko stosowane np. w medycynie, ekonomii, analizie koszykowej, zwłaszcza w zastosowaniach, gdzie mamy dostęp do dużych wielowymiarowych zmiennych, które chcemy połączyć ze sobą nieznaną z góry zależnością. W naszym przypadku została ona wykorzystana tak aby wprowadzając dane wejściowe, które sieć jeszcze nie widziała, jest w stanie uaktywnić neuron, odpowiedzialny za daną klasę i w efekcie wydać werdykt klasyfikujący nowotwór. Należy jednak pamiętać, że również poza pozytywnymi atutami tego typu sieci, występują mankamenty związane z regułami określenia parametrów sieci (liczba neuronów, wagi czy wymiary). Zły dobór któregoś z tych parametrów może skutkować nieodpowiednim procesem trenowania, co w efekcie da niepożądane oceny obiektów. Rozwiązaniem tego problemu może być wprowadzenie uczenia rozmytego, które jednak powoduje utratę ważnej cechy SOM, jaką jest przestrzenne uporządkowanie neuronów w warstwie konkurencyjnej pod względem podobieństwa reprezentowanych przez nie obiektów [1].

Program w języku MATLAB

```
%% Kacper Kilianek (305375), Adam Piszczek (303803)
[zespół nr. 22]
% Sieci neuronowe w zastosowaniach biomedycznych (SNB) –
Projekt
% Projekt nr. 36: Diagnostyka raka piersi w badaniach
mammograficznych za pomocą sieci SOM (katalog:
Mammographic Mass_MLR)

%% ========= Przygotowanie środowiska =========

clear;
clc;
clf;
close all;

% Jeżeli nie ma folderu na wykresy, stwórz go
if ~exist("./wykresy", 'dir')
       mkdir("./wykresy")
end

%% ========= Wczytanie danych =========

% Dane wczytane są w kolejnych kolumnach BI-
RADS,Age,Shape,Margin,Density,Severity
try
    M = readtable('mammographic_masses.data.txt'); %
przekonwertowanie plików na txt
catch
    fprintf("Nie udało się otworzyć pliku
mammographic_masses.data.txt")
end

%% ========= Preprocessing danych (etap 1) =========

size1 = size(M,1); % zapisanie ilości wektorów cech
przed preprocessingiem
M = table2array(M); % zamiana na dane numeryczne

% Proponowany sposób kodowania danych numerycznych ->
single
% (single-precision number), ponieważ mamy do czynienia
z liczbami
% całkowitymi typu integer
single(M);

% usunięcie złych danych pierwszej cechy (BI-RADS) z
poza zdefiniowanego zakresu 1-5
% ustawienie warunków pilnujących podane zakresy cech:

% Pierwszy wektor cech (BI-RADS):
Cond1 = M(:,1) > 5;
Cond2 = M(:,1) < 1;
Condition1 = Cond1 | Cond2; % połaczenie warunków

% Trzeci wektor cech (Kształt):
Cond1 = M(:,3) > 4;
Cond2 = M(:,3) < 1;
Condition2 = Cond1 | Cond2;

% Czwarty wektor cech (Margines):
Cond1 = M(:,4) > 5;
Cond2 = M(:,4) < 1;
Condition3 = Cond1 | Cond2;

% Trzeci wektor cech (Gęstość):
Cond1 = M(:,5) > 4;
Cond2 = M(:,5) < 1;
Condition4 = Cond1 | Cond2;

Conditions = Condition1 | Condition2 | Condition3 |
Condition4;
M(Conditions,:) = [];
size2 = size(M,1); % wielkość macierzy po redukcji
błędnych wektorów
numOfDeletedRows = size1 - size2; % ilość danych, które
zostały usunięte

%% ========= Preprocessing danych (etap 2) =========

% Podział danych na złośliwe i łagodne, tak aby
przedstawić histogramy cech
% w klasach, które są klasyfikowane
M = sortrows(M,6);
malignant = M(:,6) == 1;
benign = M(:,6) == 0;
Malignant = M(malignant,:); % zbiór cech dla przypadku
nowotworu złośliwego
Benign = M(benign,:); % zbiór cech dla przypadku
nowotworu łagodnego

%% ========= Odchylenia standardowe i średnie =========
```

```
%Parametry BI-RADS
bsr=mean(M(:,1),'omitnan');
bbsr=mean(Benign(:,1),'omitnan');
bmsr=mean(Malignant(:,1),'omitnan');
bstd=std(M(:,1),'omitnan');
bbstd=std(Benign(:,1),'omitnan');
bmstd=std(Malignant(:,1),'omitnan');
%Parametry wieku
asr=mean(M(:,2),'omitnan');
absr=mean(Benign(:,2),'omitnan');
amsr=mean(Malignant(:,2),'omitnan');
astd=std(M(:,2),'omitnan');
abstd=std(Benign(:,2),'omitnan');
amstd=std(Malignant(:,2),'omitnan');
%Parametry kształtu zmiany
ssr=mean(M(:,3),'omitnan');
sbsr=mean(Benign(:,3),'omitnan');
smsr=mean(Malignant(:,3),'omitnan');
sstd=std(M(:,3),'omitnan');
sbstd=std(Benign(:,3),'omitnan');
smstd=std(Malignant(:,3),'omitnan');
%Parametry marginesu zmiany
msr=mean(M(:,4),'omitnan');
mbsr=mean(Benign(:,4),'omitnan');
mmsr=mean(Malignant(:,4),'omitnan');
mstd=std(M(:,4),'omitnan');
mbstd=std(Benign(:,4),'omitnan');
mmstd=std(Malignant(:,4),'omitnan');
%Parametry gęstości zmiany
dsr=mean(M(:,5),'omitnan');
dbsr=mean(Benign(:,5),'omitnan');
dmsr=mean(Malignant(:,5),'omitnan');
dstd=std(M(:,5),'omitnan');
dbstd=std(Benign(:,5),'omitnan');
dmstd=std(Malignant(:,5),'omitnan');
%Grupowanie zmiennych
birads=[bsr,bstd;bbsr,bbstd;bmsr,bmstd]
wiek=[asr,astd;absr,abstd;amsr,amstd]
ksztalt=[ssr,sstd;sbsr,sbstd;smsr,smstd]
margines=[msr,mstd;mbsr,mbstd;mmsr,mmstd]
gestosc=[dsr,dstd;dbsr,dbstd;dmsr,dmstd]
%% ========= Histogramy =========

figure(1)
axis tight
hold on
subplot(2,1,1)
histogram(Benign(:,1))
title('Histogramy wartości BI-RADS dla klasy łagodnej')
subplot(2,1,2)
histogram(Malignant(:,1),'FaceColor','#A2142F')
title('Histogramy wartości BI-RADS dla klasy złośliwej')
saveas(gcf,"./wykresy/histogramy_birads.png");
hold off
figure(2)
hold on
title('Histogramy wartości BI-RADS dla obu klas')
histogram(Benign(:,1))
histogram(Malignant(:,1),'FaceColor','#A2142F')
legend('Zmiana łagodna','Zmiana złośliwa')
saveas(gcf,"./wykresy/oba_histogramy_birads.png");
hold off


figure(3)
axis tight
hold on
subplot(2,1,1)
histogram(Benign(:,2))
title('Histogramy rozkładu wieku badanych dla klasy
łagodnej')
subplot(2,1,2)
histogram(Malignant(:,2),'FaceColor','#A2142F')
title('Histogramy rozkładu wieku badanych dla klasy
złośliwej')
saveas(gcf,"./wykresy/histogramy_age.png");
hold off
figure(4)
hold on
title('Histogramy rozkładu wieku badanych dla obu klas')
histogram(Benign(:,2))
histogram(Malignant(:,2),'FaceColor','#A2142F')
legend('Zmiana łagodna','Zmiana złośliwa')
saveas(gcf,"./wykresy/oba_histogramy_age.png");
hold off


figure(5)
axis tight
hold on
subplot(2,1,1)
histogram(Benign(:,3))
title('Histogramy kształtu masy dla klasy łagodnej')
subplot(2,1,2)
histogram(Malignant(:,3),'FaceColor','#A2142F')
title('Histogramy kształtu masy dla klasy złośliwej')
saveas(gcf,"./wykresy/histogramy_shape.png");
hold off
figure(6)
hold on
title('Histogramy kształtu masy dla obu klas')
histogram(Benign(:,3))
histogram(Malignant(:,3),'FaceColor','#A2142F')
legend('Zmiana łagodna','Zmiana złośliwa')
saveas(gcf,"./wykresy/oba_histogramy_shape.png");
hold off


figure(7)
axis tight
hold on
subplot(2,1,1)
histogram(Benign(:,4))
title('Histogramy marginesu masy dla klasy łagodnej')
subplot(2,1,2)
histogram(Malignant(:,4),'FaceColor','#A2142F')
title('Histogramy marginesu masy dla klasy złośliwej')
saveas(gcf,"./wykresy/histogramy_margin.png");
hold off
figure(8)
hold on
title('Histogramy marginesu masy dla obu klas')
histogram(Benign(:,4))
histogram(Malignant(:,4),'FaceColor','#A2142F')
legend('Zmiana łagodna','Zmiana złośliwa')
saveas(gcf,"./wykresy/oba_histogramy_margin.png");
hold off


figure(9)
axis tight
hold on
subplot(2,1,1)
histogram(Benign(:,5))
title('Histogramy gęstości masy dla klasy łagodnej')
subplot(2,1,2)
histogram(Malignant(:,5),'FaceColor','#A2142F')
title('Histogramy gęstości masy dla klasy złośliwej')
saveas(gcf,"./wykresy/histogramy_density.png");
hold off
figure(10)
hold on
title('Histogramy gęstości masy dla obu klas')
histogram(Benign(:,5))
histogram(Malignant(:,5),'FaceColor','#A2142F')
legend('Zmiana łagodna','Zmiana złośliwa')
saveas(gcf,"./wykresy/oba_histogramy_density.png");
hold off

figure(11)
histogram(M(:,6))
axis tight
title("Histogram przewlekłości masy")
xlabel("Przewlekłość (0 - łagodna, 1 - złośliwa)")
ylabel("Ilość")
saveas(gcf,"./wykresy/histogram_severity.png");
```

```
%% ========= Parametry statystyczne wartości cech
=========

format long % zmiana wyświetlania dokładności zmiennych
numerycznych

% wyświetlenie wartości średnich cech
wartoscsredniabirads = mean(M(:,1));
wartoscsredniawieku = mean(M(:,2));
wartoscsredniakształtu = mean(M(:,3));
wartoscsredniamarginesumasy = mean(M(:,4));
wartoscsredniagestoscimasy = mean(M(:,5));
wartoscsredniaprzewleklosci = mean(M(:,6));

% wyświetlenie odchyleń standardowych cech
odchyleniestandardowebirads = std(M(:,1));
odchyleniestandardowewieku = std(M(:,2));
odchyleniestandardowekształtu = std(M(:,3));
odchyleniestandardowemarginesumasy = std(M(:,4));
odchyleniestandardowegestoscimasy = std(M(:,5));
odchyleniestandardoweprzewleklosci = std(M(:,6));

%% ========= Preprocessing danych (etap 3) =========

% Przed podaniem danych wejściowych do sieci neuronowej
musimy jeszcze
% przeprowadzić operację normalizacji, tak aby każda z
cech miała
% identyczny wpływ na proces uczenia się sieci. Jeżeli
nie dokonalibyśmy
% takich operacji, to jedna z cech (w naszym przypadku
dotycząca wieku),
% miałaby największy wpływ na rozkład danych, przez to,
że jej rozpiętość
% jest największa wynosząca od 18 do 96 lat. Dlatego też
poniżej
% dokonaliśmy przekształceń tak aby każda cecha
przyjmowała wartości w
% zakresie od 0 do 1.

% Zgodnie ze wzorem źródło:
http://lh3.ggpht.com/_MrdHIr826C4/Sl7eVpJOYfI/AAAAAAAAB3
4/MCFvz1r_CZQ/s800/7.JPG
M(:,1) = (M(:,1)-min(M(:,1)))/(max(M(:,1)-min(M(:,1))))
* (1-0) + 0;
M(:,2) = (M(:,2)-min(M(:,2)))/(max(M(:,2)-min(M(:,2))))
* (1-0) + 0;
M(:,3) = (M(:,3)-min(M(:,3)))/(max(M(:,3)-min(M(:,3))))
* (1-0) + 0;
M(:,4) = (M(:,4)-min(M(:,4)))/(max(M(:,4)-min(M(:,4))))
* (1-0) + 0;
M(:,5) = (M(:,5)-min(M(:,5)))/(max(M(:,5)-min(M(:,5))))
* (1-0) + 0;

%% ========= Podział danych na zbiór uczący i testowy
=========

Test = [Malignant(1:uint64(size(Malignant,1)/2),:) ;
Benign(1:uint64(size(Benign,1)/2),:)]; % dane testowe
Train =
[Malignant(uint64(size(Malignant,1)/2)+1:size(Malignant,
1),:) ;
Benign(uint64(size(Benign,1)/2)+1:size(Benign,1),:)]; %
dane uczące
```

Spis ilustracji i tabel

Rysunki

Tabele

Bibliografia

[1] K. Bartos, "SIEĆ SOM JAKO PRZYKŁAD SIECI SAMOORGANIZUJĄCEJ SIĘ," in *Econometrics. Ekonometria. Advances in Applied Data Analytics*, Wrocław.

[2] Wikipedia, "BI-RADS," [Online]. Available: https://pl.wikipedia.org/wiki/BI-RADS.

[3] P. Mazurek, "Materiały wykładowe z przedmiotu Sieci neuronowe w zastosowaniach biomedycznych," Warszawa.

[4] Wikipedia, "Self organizing map," [Online]. Available: https://en.wikipedia.org/wiki/Self-organizing_map. [Accessed 14 Kwiecień 2022].

[5] Data Mining Blog, "Data Preprocessing – Normalization," [Online]. Available: http://intelligencemining.blogspot.com/2009/07/data-preprocessing-normalization.html. [Accessed 14 Kwiecień 2022].

[6] Mathworks, [Online]. Available: https://www.mathworks.com/.

[7] J.-W. Ahn and S. Y. Syn, "Self-Organizing Maps," [Online]. Available: https://sites.pitt.edu/~is2470pb/Spring05/FinalProjects/Group1a/tutorial/som.html.

[8] M. Elter, R. Schulz-Wendtland and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process".*Medical Physics.*