Przewidywanie przewlekłej niewydolności nerek za pomocą sieci MLP

Mikołaj Pyka

1. Wstęp

Przewlekła niewydolność nerek jest pogłębiającym się i nieodwracalnym schorzeniem, które hamuje pracę nerek. Nerki spełniają w organizmie wiele funkcji, między innymi:

- dbają o prawidłowy stan płynów ustrojowych
- odpowiadają za wydalanie z organizmu szkodliwych produktów przemiany materii
- filtrują i zatrzymują składniki przydatne do codziennego funkcjonowania.

Mają również duże znaczenie dla układu kostnego oraz prawidłowej równowagi kwasowo - zasadowej krwi i na jej ciśnienie tętnicze. Oznacza to, że w przypadku uszkodzenia nerek w organizmie zaczną zbierać się szkodliwe substancje i jest on bardziej podatny na uciążliwe choroby.

Badaniami stosowanymi w diagnostyce przewlekłej niewydolności nerek są badania krwi oraz moczu wykonywane na czczo. Między innymi sprawdza się stężenie sodu i potasu, bada się czy pacjent nie ma obniżonej liczby erytrocytów oraz oblicza się współczynnik przesączania kłębuszkowego na podstawie stężenia kreatyniny. Poza tym bada się stężenie mocznika i kreatyniny oraz bada się zdolność zagęszczania moczu.

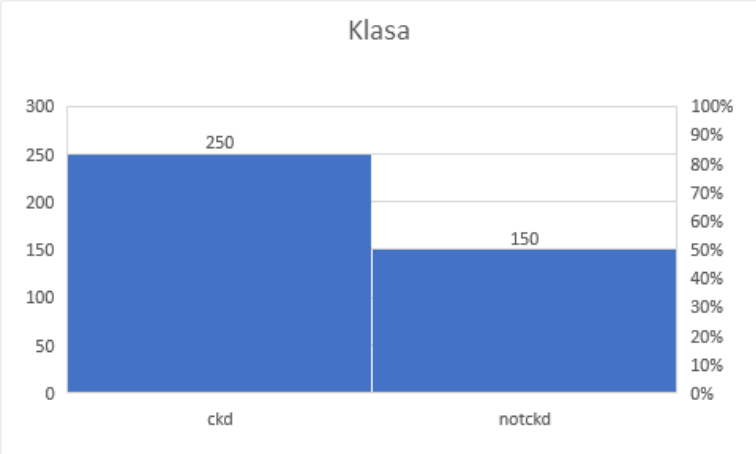
2. Dane

Dane, które będą przetwarzane zawierają dwadzieścia pięć cech w czterystu różnych instancjach. Poniżej znajduje się tabela z nazwami cech oraz innymi istotnymi parametrami ich dotyczącymi.

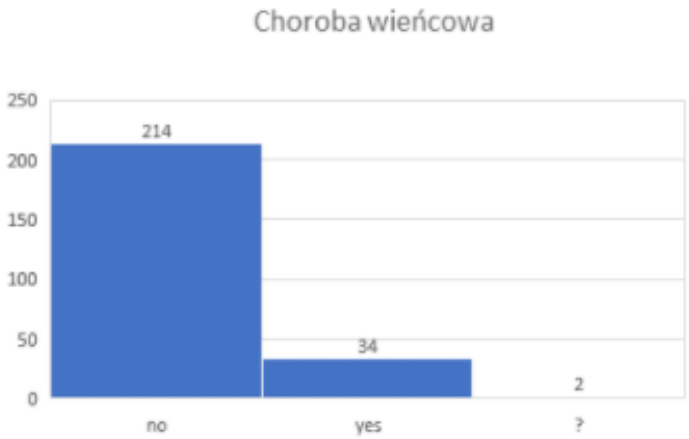
Nazwa	Typ	Jednostki (numeryczne)/ reprezentacja (nominalne)	Średnia dla zdrowych	Średnia dla chorych	Odchylenie standardowe dla zdrowych	Odchylenie standardowe dla chorych	Liczba niekompletnych danych
Wiek	Numeryczna	Lata	46,5	54,5	15,6	17,4	9
Ciśnienie krwi	Numeryczna	mm Hg	71,4	79,6	8,5	15,2	12
Gęstość względna	Nominalna	1.005,1.010,1.015,1.020,1.025	-	-			47
Albumina	Nominalna	0, 1, 2, 3, 4, 5	-	-			46
Cukier	Nominalna	0, 1, 2, 3, 4, 5	-	-			49
Czerwone krwinki	Nominalna	W normie, poza normą	-	-			152
Komórki ropnia	Nominalna	W normie, poza normą	-	-			65
Grudki komórek	Nominalna	Obecność, brak obecności	-	-			4
Bakterie	Nominalna	Obecność, brak obecności	-	-			4
Poziom glukozy	Numeryczna	mgs/dl	107,7	175,4	18,5	91,9	44
Mocznik we krwi	Numeryczna	mgs/dl	32,8	72,4	11,4	58,5	19
Kreatynina w surowicy	Numeryczna	mgs/dl	0,9	4,4	0,3	6,9	17
Sód	Numeryczna	mEq/L	141,7	133,9	4,8	12,4	87
Potas	Numeryczna	mEq/L	4,3	4,9	0,6	4,3	88
Hemoglobina	Numeryczna	gms	15,2	10,6	1,3	2,2	52
Hematokryt	Numeryczna	Brak	46,3	32,8	4,1	7,2	70
Liczba białych krwinek	Numeryczna	cells/cumm	7705,6	9093,3	1833,3	3584,3	105
Liczba czerwonych krwinek	Numeryczna	millions/cmm	5,4	3,9	0,6	0,9	130
Nadciśnienie	Nominalna	tak, nie	-	-			2
Cukrzyca	Nominalna	tak, nie	-	-			2
Choroba wieńcowa	Nominalna	tak, nie	-	-			2
Apetyt	Nominalna	dobry, słaby	-	-			1
Obrzęk stóp	Nominalna	tak, nie	-	-			1
Anemia	Nominalna	tak, nie	-	-			1
Klasa (zdrowy lub chory)	Nominalna	ckd, notckd	-	-			0

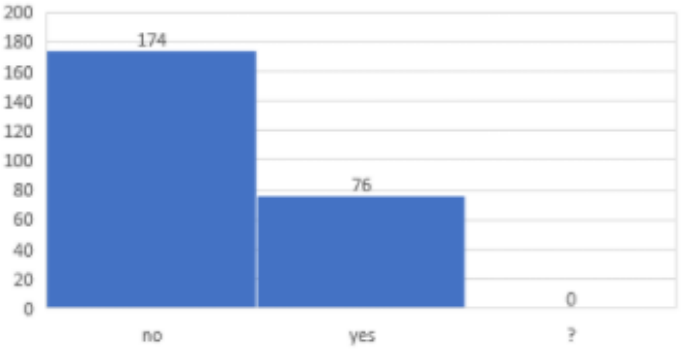
dl - decylitr, L - litr, Eq -gramorównoważnik, L - litr, gms - gramy, mgs - miligramy, mm Hg - milimetr słupa rtęccumm - milimetr sześcienny, millions/cmm - miliony na mikrolitr.

Histogramy cech nienumerycznych *(? oznacza brak danych)*:

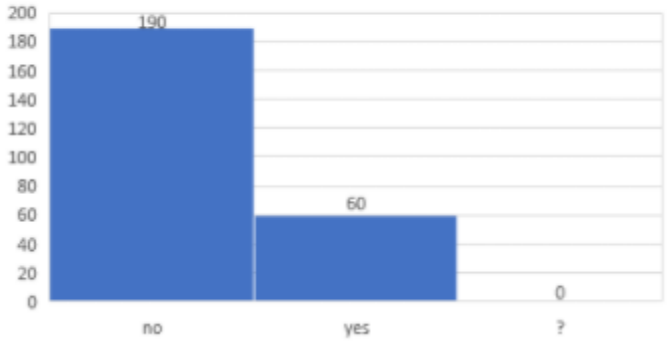


Dla chorych:

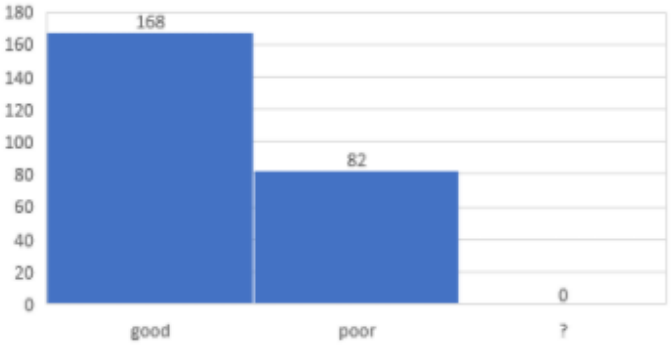




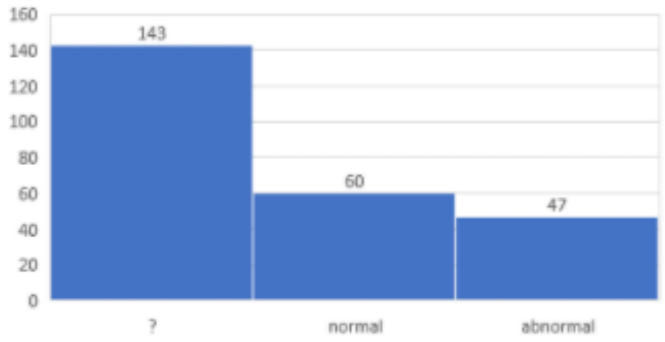
Anemia



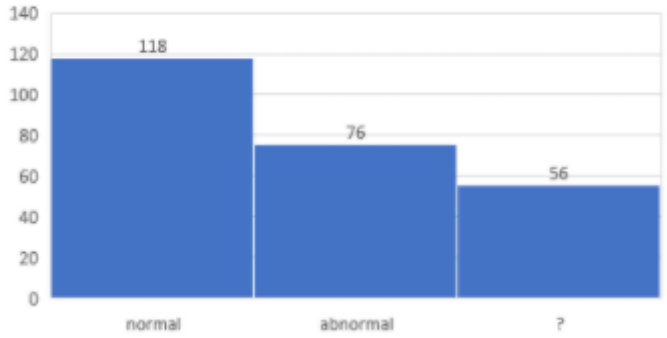
Apetyt



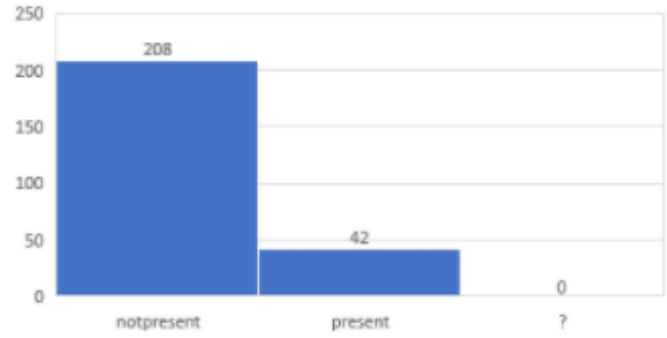
Czerwone krwinki



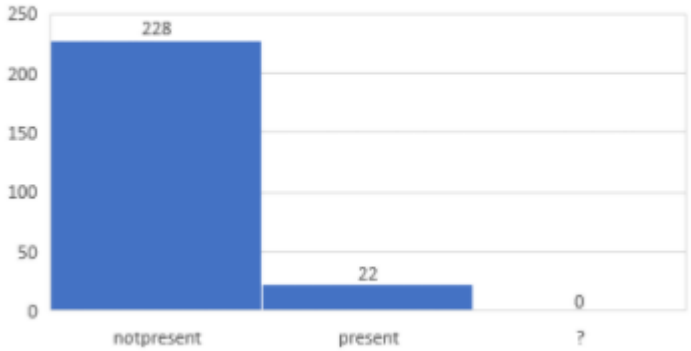
Komórki ropnia



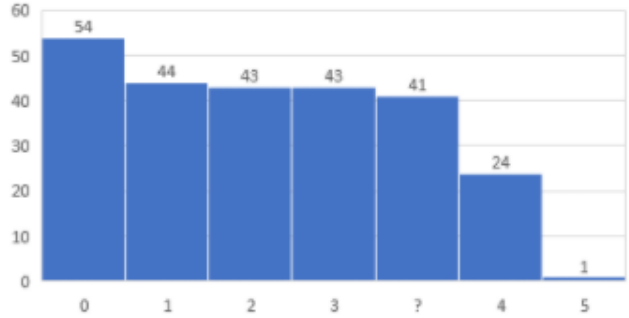
Grudki komórek ropnych



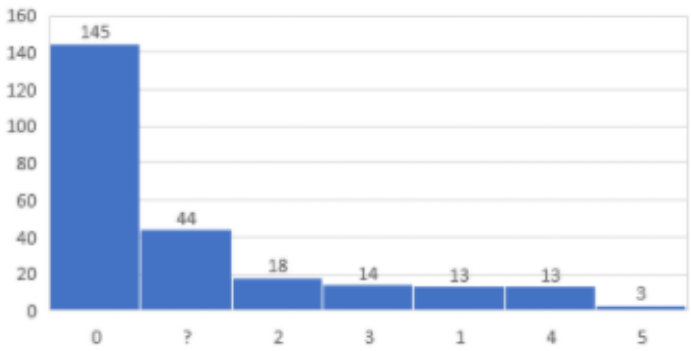
Bakterie

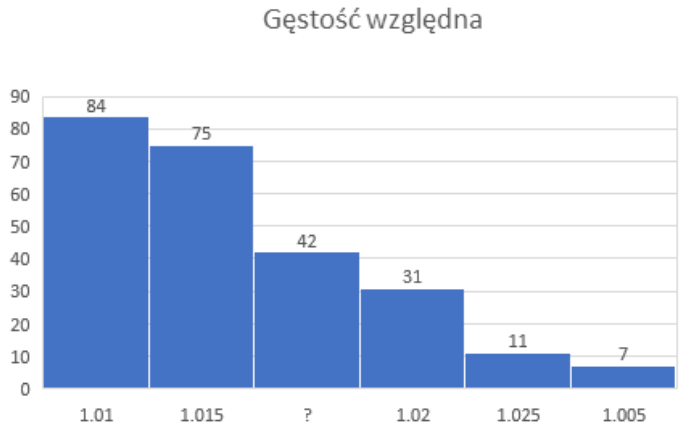


Albumina

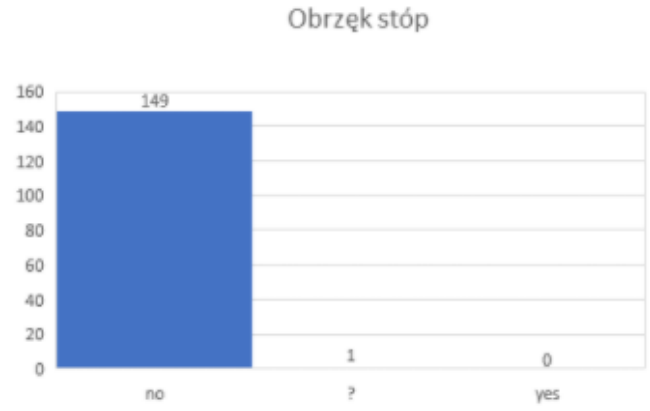


Cukier

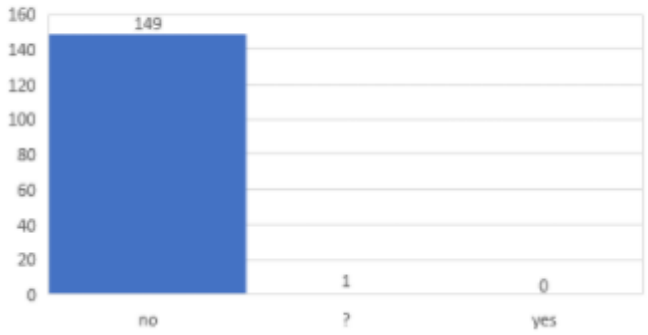




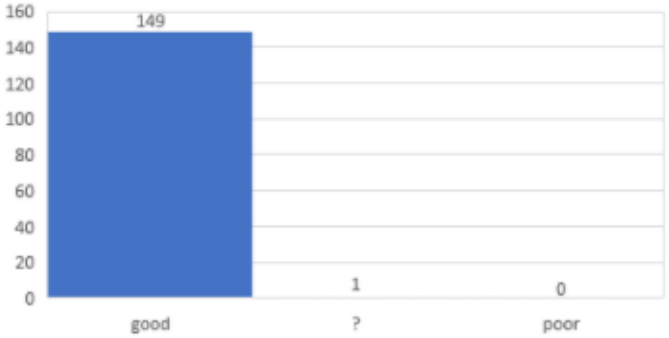
Dla zdrowych:



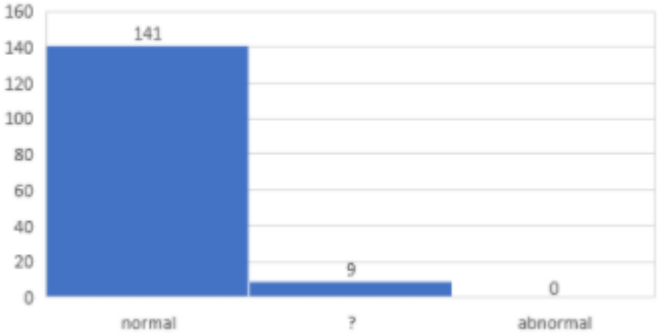
Anemia



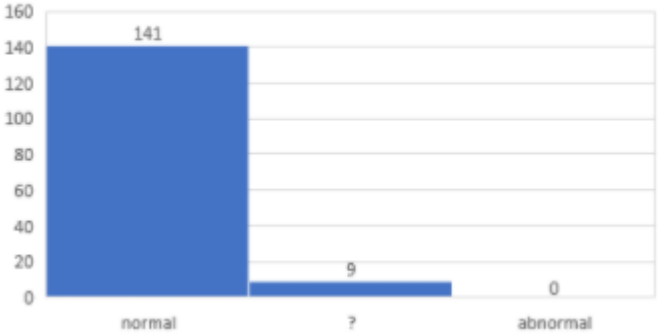
Apetyt



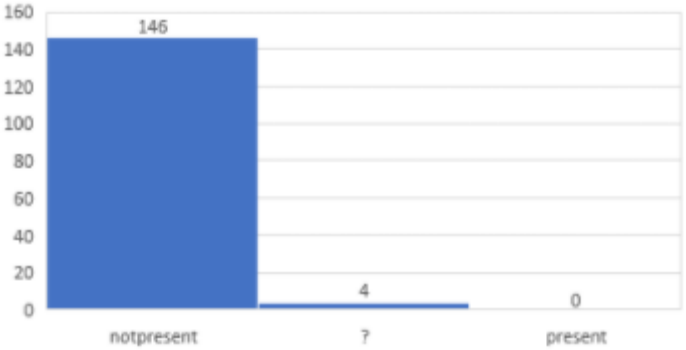
Czerwone krwinki



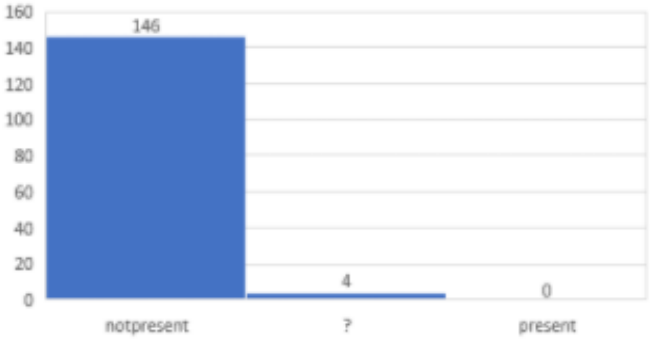
Komórki ropnia



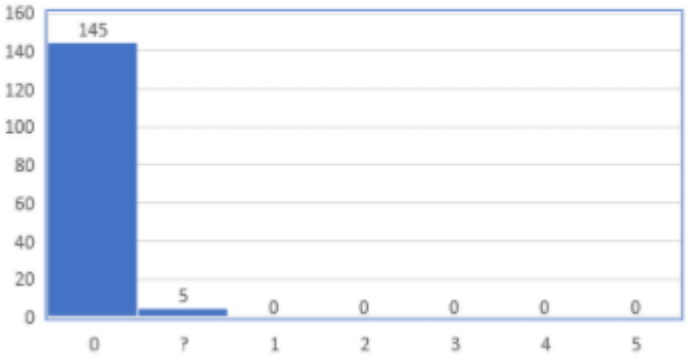
Grudki komórek ropnych



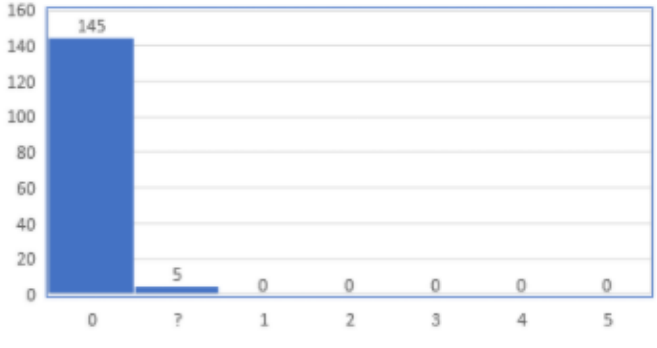
Bakterie



Albumina



Cukier



3. Przetwarzanie danych przed podaniem ich na wejście sieci

W przypadku danych niekompletnych, zastępuję je średnią wartości dla danej cechy.
Natomiast w przypadku, kiedy liczba niekompletnych danych danej cechy przekracza 25%, cechy tej nie będę brać pod uwagę. Tymi cechami są: czerwone krwinki, liczba czerwonych krwinek, liczba białych krwinek (zaznaczone w tabeli na szaro).
Dane przed podaniem ich na wejście sieci zostaną podane standaryzacji. Każdy element wejściowy zostanie przekształcony według wzoru:

$$Xn = \frac{x_N - \mu}{\sigma}$$

$Xn -$ $N - ty\ element\ wektora\ po\ stadaryzacji$
$x_N -$ $N - ty\ element\ wektora\ przed\ stadaryzacją$
$\mu -$ $wartość\ średnia$
$\sigma -$ $odchylenie\ standardowe$

4. Kodowanie danych nienumerycznych

Cechy, których dane nie są przedstawione w sposób numeryczny i które zakodowane zostaną w inny sposób niż ten dany w zbiorze (po strzałce sposób zakodowania wartości):

W wektorze wejściowym będą to:

- Czerwone krwinki - abnormal ->0, normal -> 1
- Komórki ropnia - normal-> 1, abnormal-> 0
- Grudki komórek ropnych - present -> 1, notpresent -> 0
- Bakteria - present -> 1, notpresent -> 0
- Nadciśnienie - yes ->1, no -> 0

- Cukrzyca - yes -> 1, no -> 0
- Choroba wieńcowa - yes ->1, no -> 0
- Apetyt - good -> 1, poor -> 0
- Obrzęk stóp - yes -> 1, no -> 0
- Anemia - yes ->1, no -> 0

W wektorze wyjściowym będzie to jedna wartość - obecność choroby u badanego
ckd ->1, notckd -> 0.

6. Podział na dane treningowe i testowe

Dane będą podzielone losowo w proporcjach 70% danych uczących do 30% danych testowych. Oznacza to, że z klasy zdrowych 105 wektorów będzie przeznaczonych do uczenia, a 45 do testów, a z klasy chorujących na przewlekłą niewydolność nerek do uczenia wykorzystane będzie 175 wektorów, a do testowania 75.

7. Struktura sieci

Zaimplementowana sieć neuronowa składa się z warstwy wejściowej, jednej warstwy ukrytej oraz jednej warstwy wyjściowej.

Liczbę neuronów warstwy ukrytej wyznaczyliśmy zgodnie ze wzorem:

$$N_{ukryte} = \sqrt{N_{WE} * N_{WY}}$$

N_{WE} – liczba neuronów warstwy wejściowej (21)

N_{WY} – liczba neuronów warstwy wyjściowej (1)

Poniżej dane dotyczące liczb neuronów poszczególnych warstw:

Warstwa	Liczba neuronów
Wejściowa	21
Ukryta	5
Wyjściowa	1

Użytą funkcją aktywacji jest sigmoidalna funkcja unipolarna o postaci $f(x) = \frac{1}{1+\exp(-x)}$

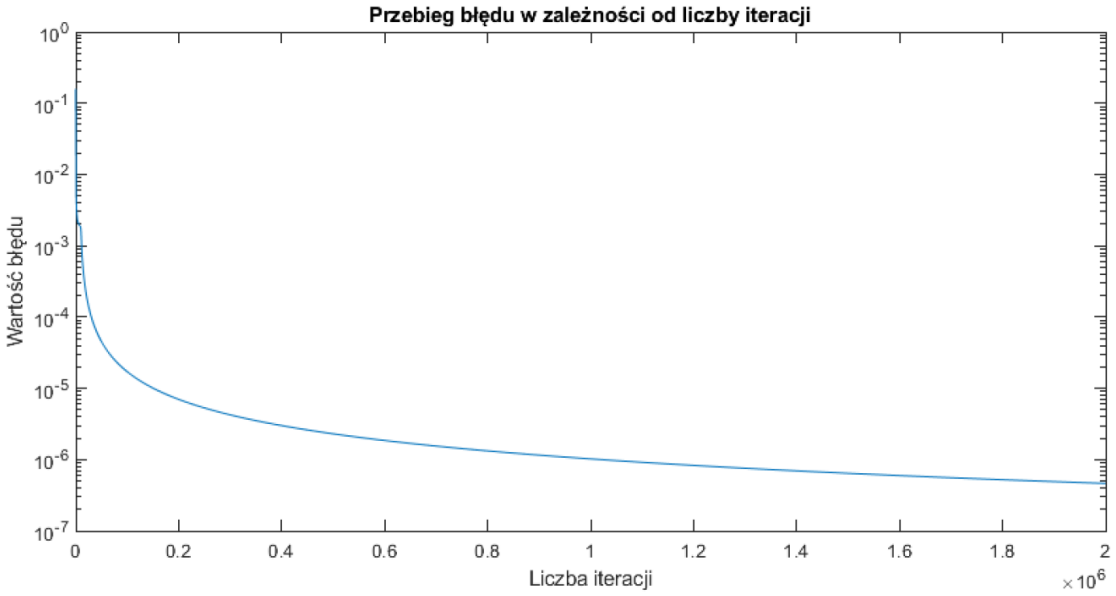
8. Szczegóły algorytmu uczenia

Do uczenia sieci zastosowany został algorytm wstecznej propagacji błędów:

1. Przyjmij losowe wartości wag warstwy ukrytej (w1) i wag warstwy wyjściowej (w2).
2. Podaj na wejście sieci dane uczące w losowej kolejności. X – macierz danych uczących.
3. Oblicz pobudzenia neuronów warstwy ukrytej $v = w1 * X$
4. Oblicz stan wyjść neuronów warstwy ukrytej $z = \frac{1}{1+\exp(-v)}$
5. Oblicz pobudzenia neuronów warstwy wyjściowej $u = w2 * z$
6. Oblicz stan wyjść neuronów warstwy wyjściowej $Y = \frac{1}{1+\exp(-u)}$
7. Oblicz sygnał błędu dla warstwy wyjściowej $\delta 2 = \frac{\partial C}{\partial Y} * f'(u),\ gdzie\ C = \frac{1}{2}(D-Y)^2$, gdzie D – oczekiwana wartość na wyjściu sieci
8. Oblicz sygnał błędu dla warstwy ukrytej $\delta 1 = f'(v) * (w2^T * \delta 2)$
9. Zmodyfikuj wagi warstwy wyjściowej $w2_{nowa} = w2 + \eta * \delta 2 * Z^T$
10. Zmodyfikuj wagi warstwy ukrytej $w1_{nowa} = w1 + \eta * \delta 1 * X^T$
11. Jeśli poziom błędu nie będzie zadowalający – powrót do punktu nr 2.

9. Przebieg błędu w zależności od liczby iteracji

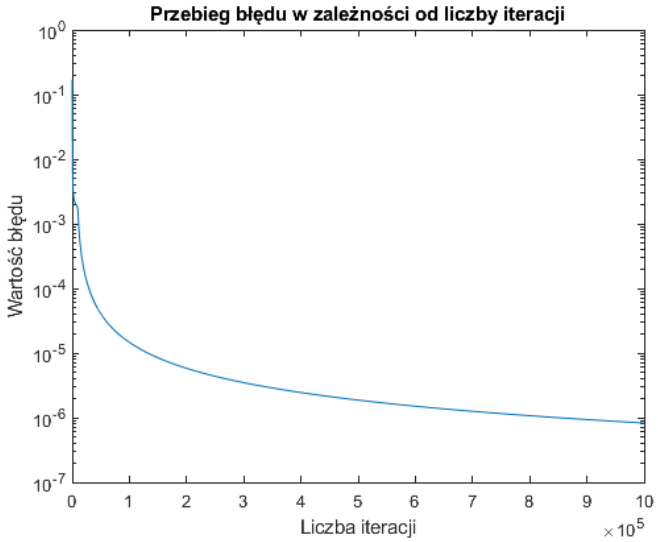
Wykonana implementacja sieci neuronowej osiągała dość zadowalające wyniki. Dla dwóch milionów iteracji czułość sieci wynosiła 93.3%, specyficzność 93.3%, a dokładność 93.3%. Przebieg błędu wyglądał następująco:



Implementacja ta zakładała obecność pięciu neuronów w warstwie ukrytej oraz wartość współczynnika uczenia $\eta = 0.02$.

Pierwotna wersja sieci miała pięć neuronów w warstwie ukrytej, a współczynnik uczenia wynosił $\eta = 0.02$. Warunkiem zakończenia trenowania sieci jest osiągnięcie błędu mniejszego niż 10^{-10} lub liczby iteracji wynoszącej 10^6.

Dla miliona iteracji, pięciu neuronów warstwy ukrytej i współczynnika uczenia $\eta = 0.02$ przebieg błędu prezentuje się następująco:



Czułość sieci wyniosła 93.3%, specyficzność 93.3%, a dokładność 93.3%.

Czas uczenia sieci wyniósł 131 sekund.

10. Czułość, specyficzność i dokładność

Dla całego zbioru istnieją konkretne ilości danych przypadków. Możliwości, które mogą nastąpić na wyjściu sieci to:

TP – prawidłowa decyzja pozytywna

TF – prawidłowa decyzja negatywna

FP – wynik fałszywie pozytywny

FN – wynik fałszywie negatywny

Na potrzeby policzenia opisanych w dalszej części miar decyzję za prawidłową uznałem, jeśli podczas testów dla wartości na wyjściu sieci (Y) oraz wartości oczekiwanych na wyjściu sieci (D) zachodzi nierówność:

$$|Y(i) - D(i)| < 0.1$$

Czułość można interpretować jako prawdopodobieństwo, że klasyfikacja będzie poprawna, pod warunkiem że przypadek jest pozytywny. Prawdopodobieństwo, że test wykonany dla osoby cierpiącej na przewlekłą niewydolność nerek wykaże że faktycznie jest ona chora.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

W przypadku mojej sieci czułość dla dwóch milionów iteracji i współczynniku uczenia $\eta = 0.02$ wyniosła 93.(3) %.

Specyficzność można interpretować jako prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem, że przypadek jest negatywny. Prawdopodobieństwo, że test wykonany dla osoby zdrowej nie wykaże że jest ona chora na przewlekłą niewydolność nerek.

$$SPC = \frac{TN}{N} = \frac{TN}{TN + FP}$$

W przypadku mojej sieci specyficzność dla dwóch milionów iteracji i współczynniku uczenia $\eta = 0.02$ wyniosła 93.(3) %.

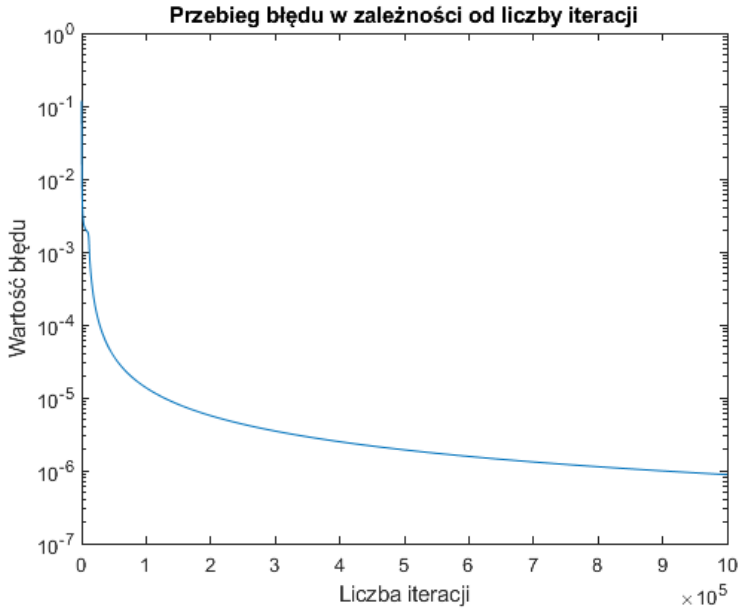
Dokładność natomiast jest prawdopodobieństwem prawidłowej klasyfikacji.

$$ACC = \frac{TP + TN}{P + N}$$

W przypadku mojej sieci dokładność dla dwóch milionów iteracji i współczynniku uczenia $\eta = 0.02$ wyniosła 93.(3) %.

11. Analiza nauki sieci po zmianie liczby neuronów w warstwie ukrytej

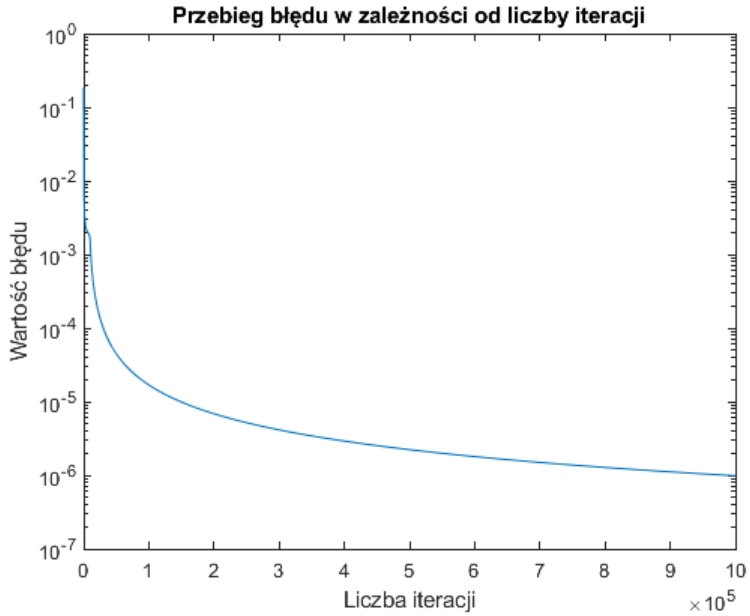
Liczba neuronów=2



Czułość sieci wyniosła 92%, specyficzność 93.3%, a dokładność 92.5%.

Czas uczenia sieci wyniósł 103 sekundy.

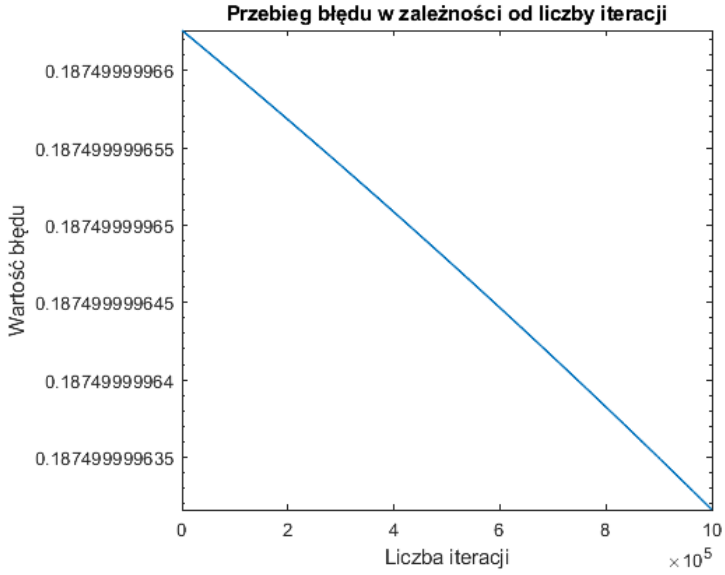
Liczba neuronów=10



Czułość sieci wyniosła 93.3%, specyficzność 93.3%, a dokładność 93.3%.

Czas uczenia sieci wyniósł 136 sekund.

Liczba neuronów=50

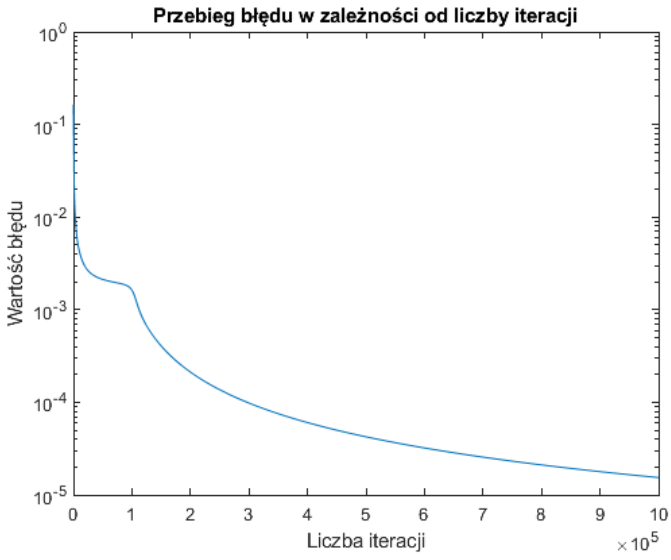


Czułość sieci wyniosła 100%, specyficzność 0%, a dokładność 62.5%.

Czas uczenia sieci wyniósł 276 sekund.

Analiza nauki sieci po zmianie współczynnika uczenia

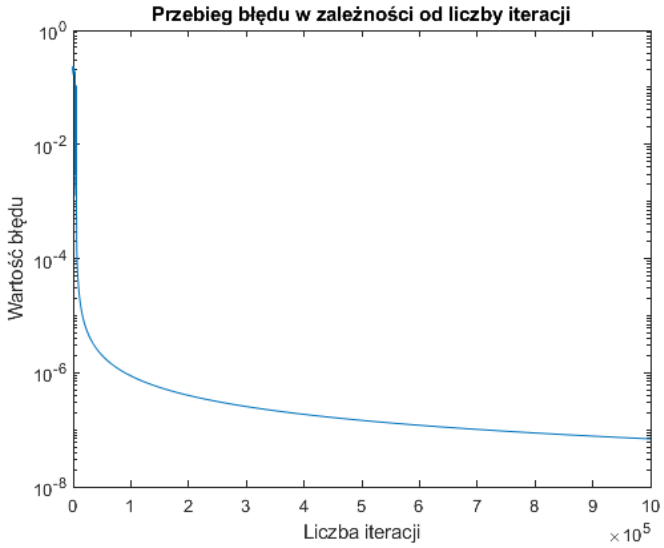
$$\eta = 0.002$$



Czułość sieci wyniosła 93.3%, specyficzność 91.1%, a dokładność 92.5%.

Czas uczenia sieci wyniósł 118 sekund.

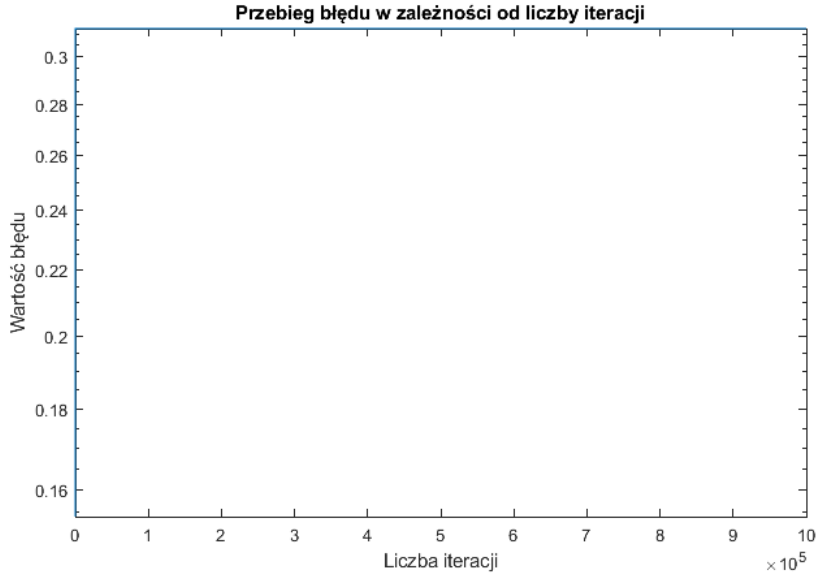
$\eta = 0.2$



Czułość sieci wyniosła 92%, specyficzność 93.3%, a dokładność 92.5%.

Czas uczenia sieci wyniósł 115 sekund.

$\eta = 2$



Czułość sieci wyniosła 0%, specyficzność 100%, a dokładność 37.5%.

Czas uczenia sieci wyniósł 115 sekund.