



Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji
Informatyczne systemy automatyki

Sieci neuronowe

Klasyfikacja chorób serca

Adam Makarewicz

Spis treści

1	Wprowadzenie	2
2	Analiza danych wejściowych	2
2.1	Wiek pacjentów	3
2.2	Płeć pacjentów	3
2.3	Płeć w zależności od bazy danych	4
2.4	Wiek pacjentów w zależności od płci	4
2.5	Rodzaj bólu w klatce piersiowej w zależności od płci	4
2.6	Rodzaj bólu w klatce piersiowej w zależności od bazy danych	5
3	Sieć neuronowa binarna	5
3.1	Wykres funkcji kosztu	5
3.2	Wykresy błędu MSE	6
3.3	Wykresy błędu klasyfikacji	6
3.4	Raport	6
4	Sieć neuronowa wieloklasowa	7
4.1	Wykres funkcji kosztu	7
4.2	Wykresy błędu MSE	8
4.3	Wykresy błędu klasyfikacji	8
4.4	Raport	8

1 Wprowadzenie

Problem, który jest przedmiotem badań, dotyczy predykcji obecności choroby serca u pacjentów na podstawie zbioru cech medycznych. Celem jest stworzenie modelu, który potrafi skutecznie rozróżnić, czy pacjent cierpi na chorobę serca (występująca w postaciach 1, 2, 3, 4) czy nie (wartość 0). Problem polega na klasyfikacji pacjentów na podstawie 14 atrybutów, takich jak wiek, poziom cholesterolu, ciśnienie krwi i inne czynniki ryzyka.

W ramach realizacji projektu opracowano dwie sieci neuronowe: sieć binarną oraz sieć wieloklasową. Sieć binarna została zaprojektowana do rozwiązania problemu klasyfikacji binarnej, polegającego na rozróżnieniu pomiędzy pacjentami, u których stwierdzono obecność choroby serca (wartości 1, 2, 3, 4) a pacjentami zdrowymi (wartość 0).

W celu bardziej zaawansowanej klasyfikacji, opracowano również sieć neuronową wieloklasową, która umożliwia przypisanie pacjentów do jednej z czterech klas, odpowiadających różnym stopniom zaawansowania choroby serca (wartości 1, 2, 3, 4). Obie sieci zostały przetestowane na tych samych danych, co pozwoliło na porównanie ich skuteczności w kontekście rozwiązywania problemu klasyfikacji medycznej.

2 Analiza danych wejściowych

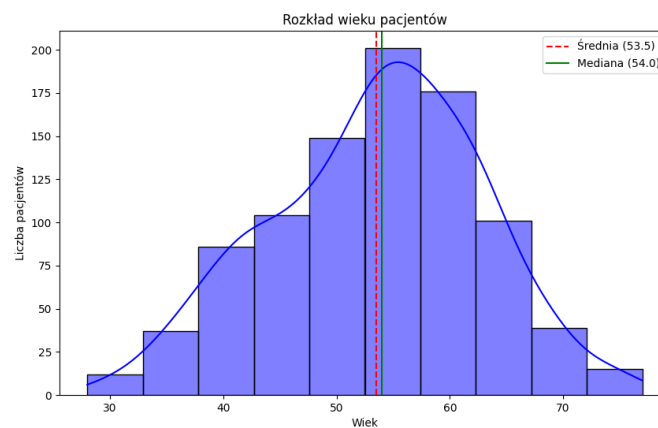
Zbiór danych zawiera atrybuty opisujące cechy pacjentów, które są istotne w analizie i diagnostyce chorób serca. Każda cecha ma swoje unikalne znaczenie i wartość w interpretacji wyników badań.

Poniżej znajduje się szczegółowy opis tych cech:

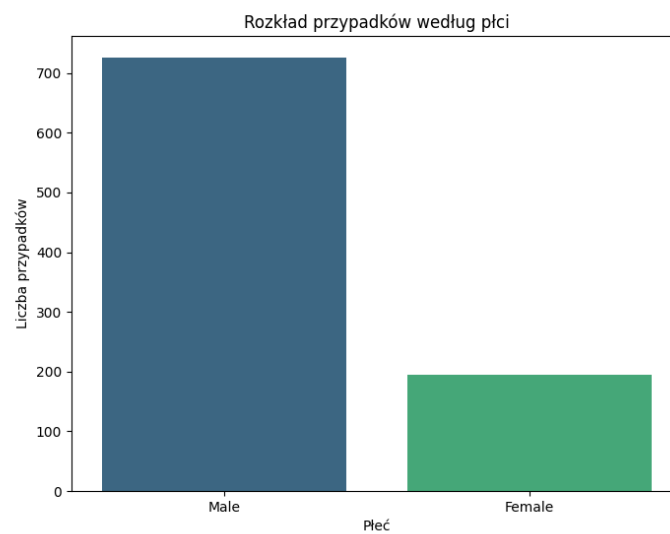
1. age – Wiek pacjenta w latach.
2. sex – Płeć pacjenta.
3. cp – Rodzaj bólu w klatce piersiowej.
4. trestbps – Ciśnienie tętnicze w spoczynku (w mm Hg) mierzone przy przyjęciu do szpitala.
5. chol – Poziom cholesterolu w surowicy krwi (w mg/dl).
6. fbs – Cukier na czczo.
7. restecg – Wynik elektrokardiogramu w spoczynku.
8. thalach – Maksymalne osiągnięte tętno podczas testu wysiłkowego.
9. exang – Wystąpienie dławicy wywołanej wysiłkiem.
10. oldpeak – Depresja odcinka ST wywołana wysiłkiem w stosunku do stanu spoczynkowego.
11. slope – Nachylenie odcinka ST w szczytowym momencie wysiłku.

12. ca – Liczba głównych naczyń krwionośnych (od 0 do 3), widocznych w badaniu fluoroskopowym.
13. thal – Wynik badania z użyciem izotopu talu.

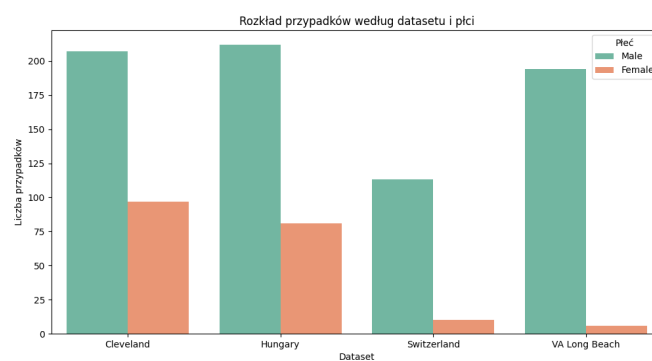
2.1 Wiek pacjentów



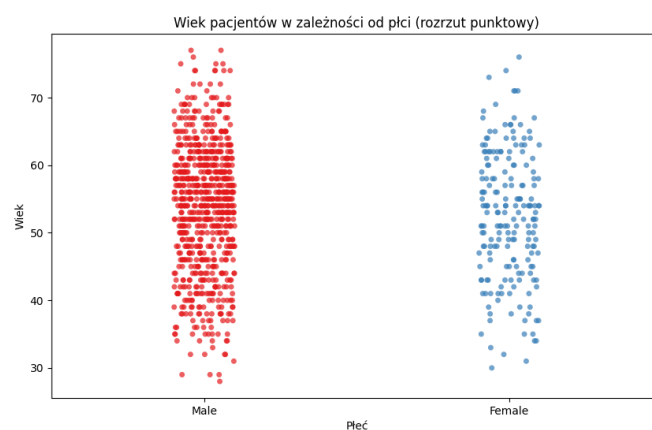
2.2 Płeć pacjentów



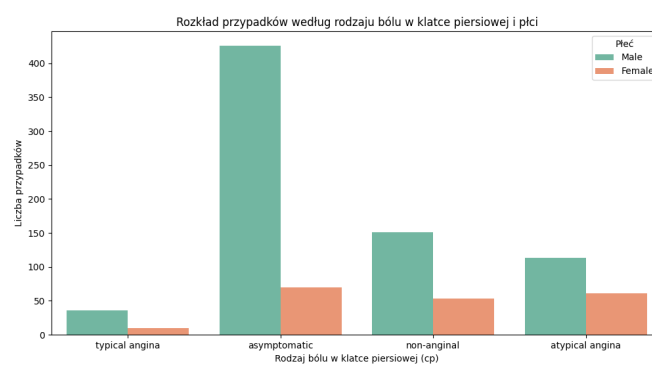
2.3 Płeć w zależności od bazy danych



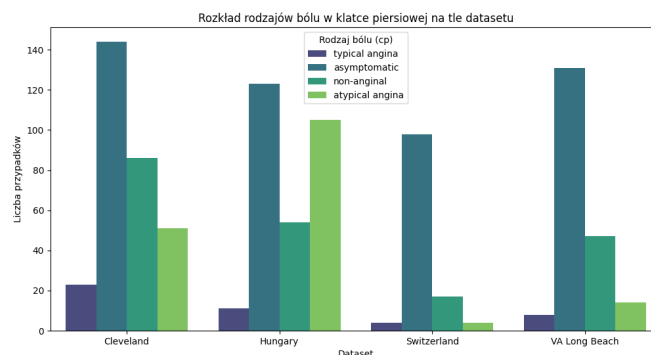
2.4 Wiek pacjentów w zależności od płci



2.5 Rodzaj bólu w klatce piersiowej w zależności od płci



2.6 Rodzaj bólu w klatce piersiowej w zależności od bazy danych



3 Sieć neuronowa binarna

W trakcie eksperymentów przeprowadzono optymalizację parametrów modelu binarnego przy użyciu metody siatki wyszukiwania (**GridSearchCV**). Model został poddany procesowi strojenia, w którym zmieniano następujące hiperparametry:

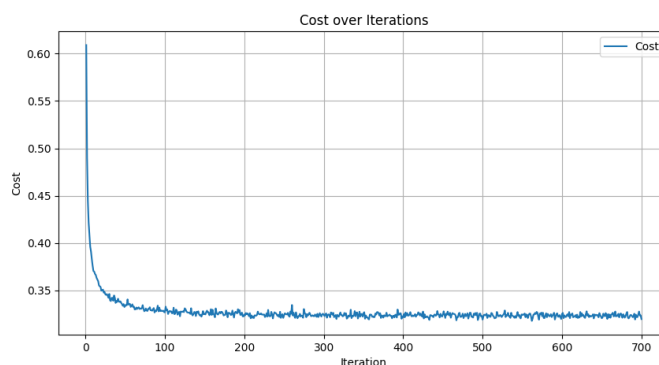
- **Liczność wsadu** (`batch_size`): liczba próbek przetwarzanych w jednym kroku uczenia,
- **Współczynnik uczenia** (`learning_rate`): tempo dostosowywania wag modelu w trakcie uczenia,
- **Liczba iteracji** (`num_iterations`): liczba pełnych przebiegów przez zbiór danych.

Na podstawie wyników procesu optymalizacji wybrano najlepsze wartości parametrów dla modelu binarnego:

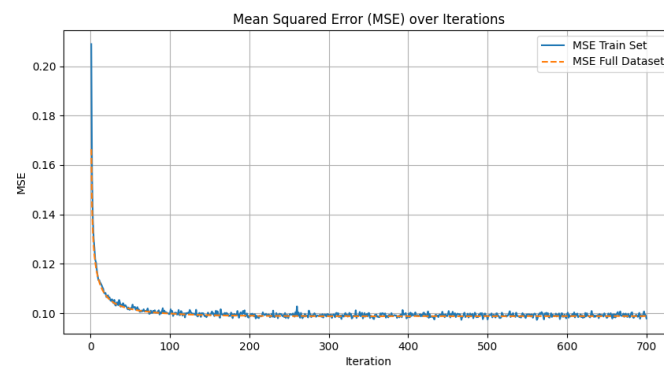
- `batch_size`: 128,
- `learning_rate`: 0.25,
- `num_iterations`: 700.

Optymalny zestaw parametrów pozwolił na uzyskanie stabilnego procesu uczenia oraz satysfakcjonujących wyników klasyfikacji w kontekście problemu.

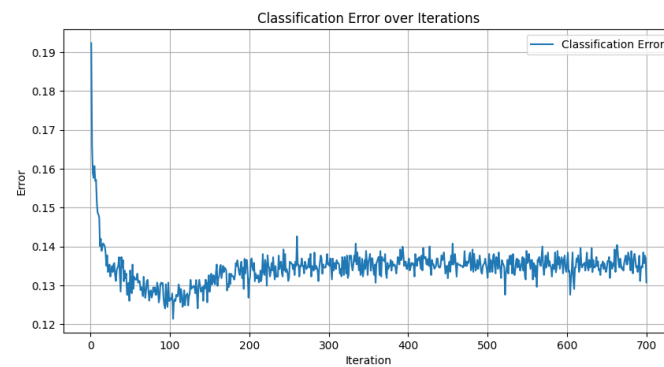
3.1 Wykres funkcji kosztu



3.2 Wykresy błędu MSE



3.3 Wykresy błędu klasyfikacji



3.4 Raport

```
Accuracy: 0.86
Confusion Matrix:
[[104  20]
 [ 24 156]]
Classification Report:
              precision    recall  f1-score   support

     0       0.81      0.84      0.83       124
     1       0.89      0.87      0.88       180

   accuracy          0.86          304
  macro avg          0.85          304
weighted avg          0.86          304

Log Loss: 0.3918
Roc auc: 0.0000
```

4 Sieć neuronowa wieloklasowa

W trakcie eksperymentów przeprowadzono również optymalizację parametrów modelu wieloklasowego, korzystając z metody siatki wyszukiwania (`GridSearchCV`). Proces strojenia obejmował następujące hiperparametry:

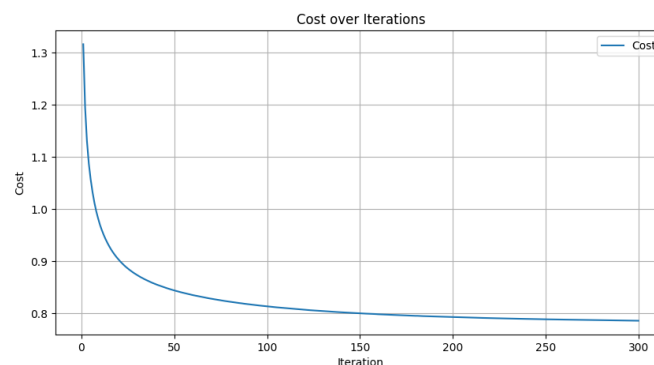
- **Liczność wsadu** (`batch_size`): liczba próbek przetwarzanych w jednym kroku uczenia,
- **Parametr β_2** (`beta2`): wykładnikowa średnia kwadratów gradientu w optymalizatorze Adam,
- **Współczynnik uczenia** (`learning_rate`): tempo dostosowywania wag modelu w trakcie uczenia,
- **Momentum** (`momentum`): dodatkowy parametr przyspieszający zbieżność uczenia,
- **Liczba iteracji** (`num_iterations`): liczba pełnych przebiegów przez zbiór danych.

Najlepsze uzyskane wartości parametrów dla modelu wieloklasowego to:

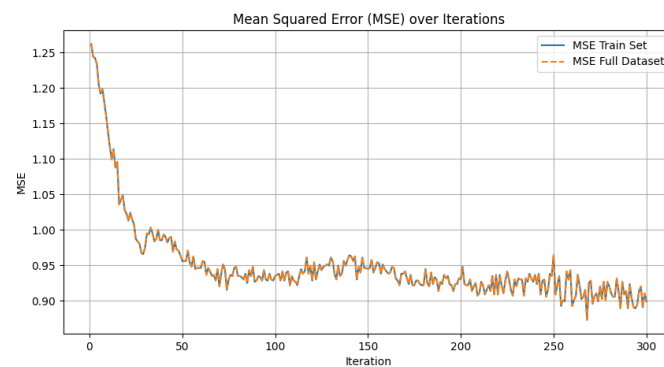
- `batch_size`: 128,
- `beta2`: 0.999,
- `learning_rate`: 0.01,
- `momentum`: 0.6,
- `num_iterations`: 300.

Optymalny zestaw parametrów pozwolił na poprawę jakości klasyfikacji w problemie wieloklasowym.

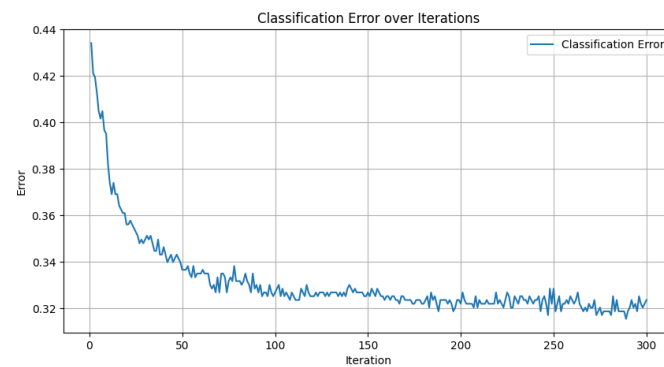
4.1 Wykres funkcji kosztu



4.2 Wykresy błędu MSE



4.3 Wykresy błędu klasyfikacji



4.4 Raport

```
Accuracy: 0.59
Confusion Matrix:
[[106  9  7  2  0]
 [ 26 55  6  6  0]
 [  4 19  6 10  4]
 [  5  7  8 13  5]
 [  0  1  1  4  0]]
Classification Report:
              precision    recall  f1-score   support

     0       0.75         0.85         0.80         124
     1       0.60         0.59         0.60          93
     2       0.21         0.14         0.17          43
     3       0.37         0.34         0.36          38
     4       0.00         0.00         0.00           6

   accuracy          0.59         304
  macro avg          0.39         0.39         0.38         304
 weighted avg          0.57         0.59         0.58         304

Log Loss: 1.0058
Roc auc: 0.8142
```