

# Sieć Kohonena

Gaspar Sekula

Maj 2025

## Spis treści

<b>1</b>	<b>Wstęp</b>	<b>2</b>
<b>2</b>	<b>Opis modelu</b>	<b>2</b>
<b>3</b>	<b>Zbiory danych</b>	<b>2</b>
<b>4</b>	<b>Eksperymenty</b>	<b>3</b>
4.1	Hexagon . . . . .	3
4.1.1	Przykładowa sieć . . . . .	3
4.1.2	Eksperymenty i wyniki . . . . .	5
4.1.3	Wizualizacje najlepszej sieci . . . . .	6
4.1.4	Wnioski . . . . .	9
4.2	Cube . . . . .	9
4.2.1	Przykładowa sieć . . . . .	9
4.2.2	Eksperymenty i wyniki . . . . .	9
4.2.3	Wizualizacje najlepszej sieci . . . . .	11
4.2.4	Wnioski . . . . .	12
4.3	MNIST . . . . .	12
4.3.1	Wybór najlepszej sieci . . . . .	12
4.3.2	Wizualizacje najlepszej sieci . . . . .	13
4.3.3	Wnioski . . . . .	15
4.4	HARUS . . . . .	15
4.4.1	Wybór najlepszej sieci . . . . .	15
4.4.2	Wizualizacje najlepszej sieci . . . . .	15
4.4.3	Wnioski . . . . .	18
<b>5</b>	<b>Podsumowanie</b>	<b>18</b>

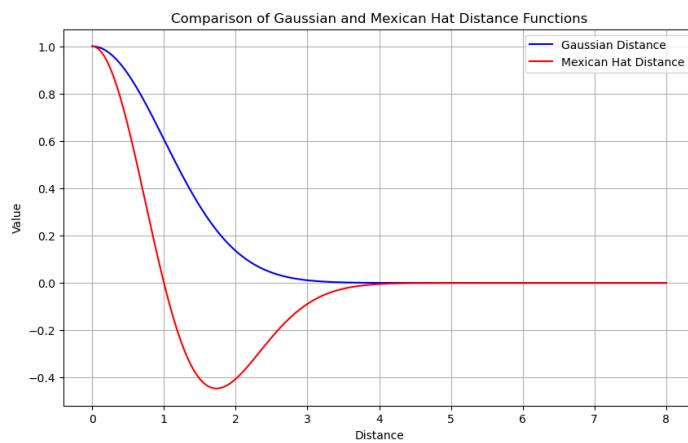
# 1 Wstęp

Raport stanowi zebranie wniosków z projektu, który polegał na implementacji sieci Kohonena. Sieć Kohonena (ang. Self Organising Map) jest to nienadzorowana technika uczenia maszynowego, która służy do tworzenia dwuwymiarowej reprezentacji wielowymiarowego zbioru danych przy zachowaniu topologicznej struktury danych. Zaimplementowano model i przeprowadzono doświadczenia na czterech zbiorach danych - dwóch prostych i dwóch bardziej skomplikowanych.

## 2 Opis modelu

Zaimplementowano model sieci Kohonena, w którym użytkownik może wybrać:

- wymiary sieci: długość i szerokość,
- rodzaj sieci: siatka kwadratowa lub sześciokątna,
- liczbę epok treningu,
- współczynnik sąsiedztwa (*neighbourhood scaler*),
- szybkość uczenia (*learning rate*),
- funkcję odległości: Gausowską (*GaussianDistance*) lub Meksykański Kapelusz (*MexicanHatDistance*) (zob. Rysunek 1).



Rysunek 1: Porównanie funkcji sąsiedztwa.

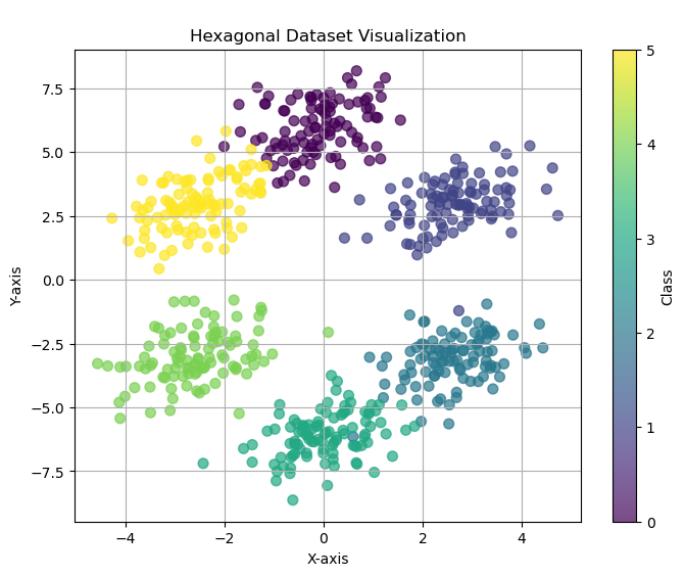
## 3 Zbiory danych

Przeanalizowano 4 zbiory danych:

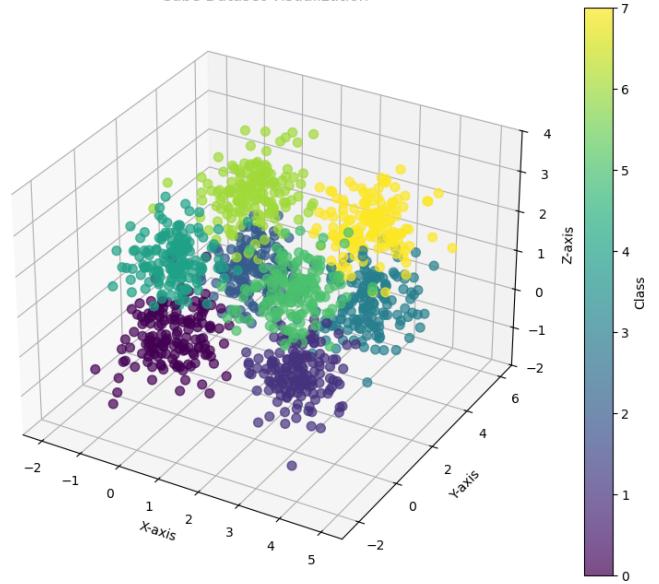
- **Hexagon**: dane dwuwymiarowe przedstawiające wierzchołki sześciokąta (Rysunek 2),
- **Cube**: dane trójwymiarowe przedstawiające wierzchołki sześcianu (Rysunek 3),
- **MNIST**: dane zawierające zestaw cyfr zapisanych ręcznie, 784 piksele (zob. Rysunek 4),
- **Human Activity Recognition Using Smartphones (HARUS)**: dane przedstawiające aktywności wykonywane przez osoby z zamontowanym smartfonem wyposażonym w czujniki inercyjne, 561 cech (zob. Rysunek 5).

Ze względu na wielowymiarowość danych MNIST i HARUS, zastosowano reprezentację dwukomponentową t-SNE. Z wizualizacji widać, że klasteryzacja zbiorów Hexagon i Cube będzie istotnie łatwiejsza niż w przypadku dwóch pozostałych, bowiem klastry są łatwo separowalne.

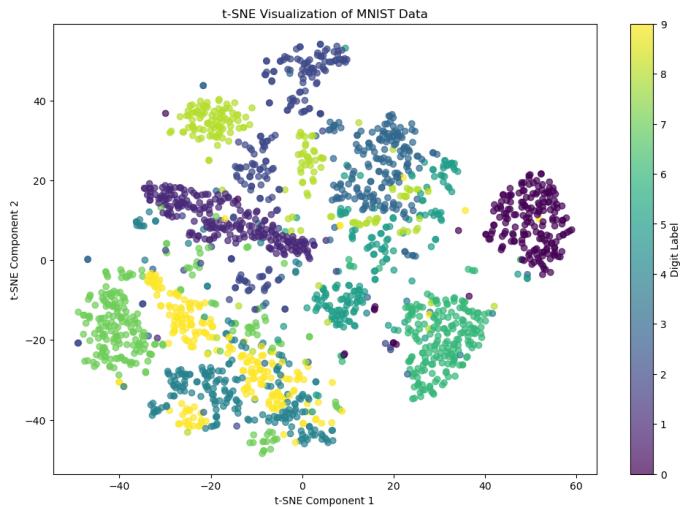
Cube Dataset Visualization



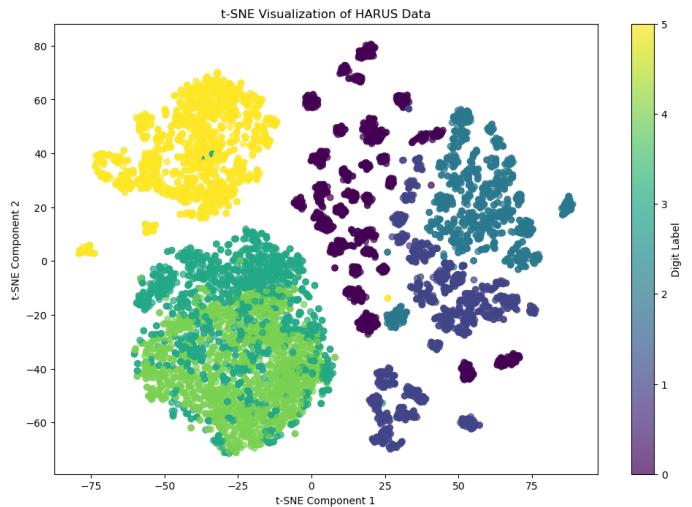
Rysunek 2: Dane Hexagon.



Rysunek 3: Dane Cube.



Rysunek 4: Dane MNIST.



Rysunek 5: Dane HARUS.

## 4 Eksperymenty

Zbadano wpływ:

- funkcji odległości (Gaussian lub MexicanHat),
- wyboru rodzaju sieci (kwadratowa lub sześciokątna),
- szybkości uczenia (0.001, 0.01, 0.1 lub 1),
- współczynnika sąsiedztwa (0.1, 0.5 lub 1)

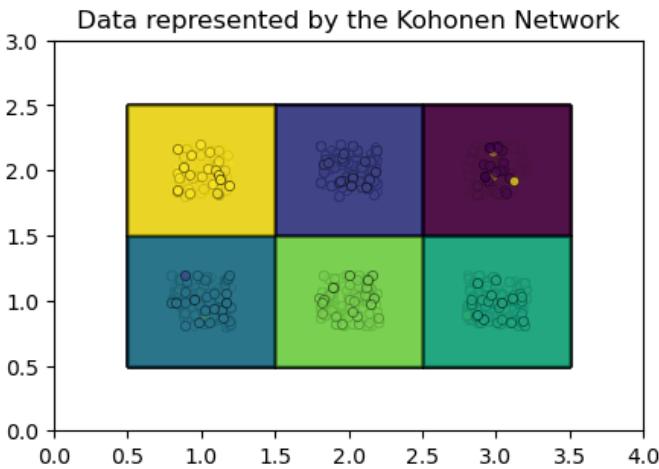
na jakość klasteryzacji mierzoną indeksem Silhouette. W doświadczeniu trening trwał 10 epok. Spośród wszystkich wyników wybrano 3 najlepsze konfiguracje (oznaczone kolorem w tabelach podsumowujących badania) i każdą z nich trenowano 5 razy przez 30 epok. Zebrane średnie i odchylenia standardowe indeksów Silhouette i Davies-Bouldin.

### 4.1 Hexagon

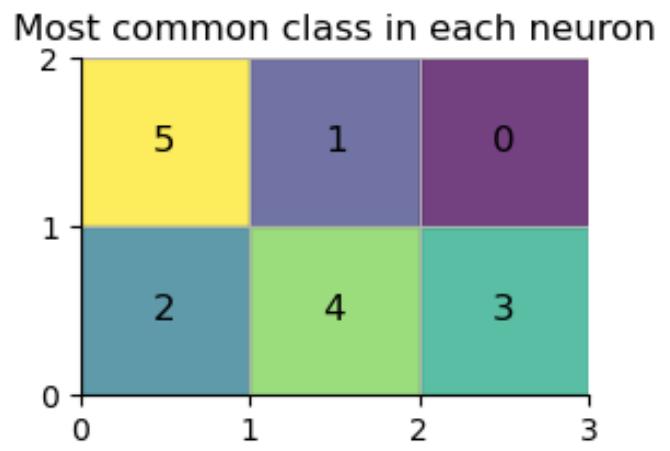
#### 4.1.1 Przykładowa sieć

Na początek wybrano sieć 6-neuronową na planie kwadratowym. Zbadano różne współczynniki dla Gausowskiej funkcji sąsiedztwa (0.1, 1, 5, 10). Wybrano najlepszą architekturę: ze współczynnikiem sąsiedztwa 0.1.

Rysunek 8 przedstawia postęp uczenia się sieci. Widać, że uzyskana klasteryzacja jest dobra: każdy neuron otacza się punktami jednej klasy. Wartość indeksu Silhouette wynosi 0.5989, a indeksu Daviesa-Bouldina - 0.5403. Na rysunku 9 można przeanalizować wyniki klasteryzacji. Decyzje o przypisaniu etykiety podjęto na podstawie najbliższego neuronu

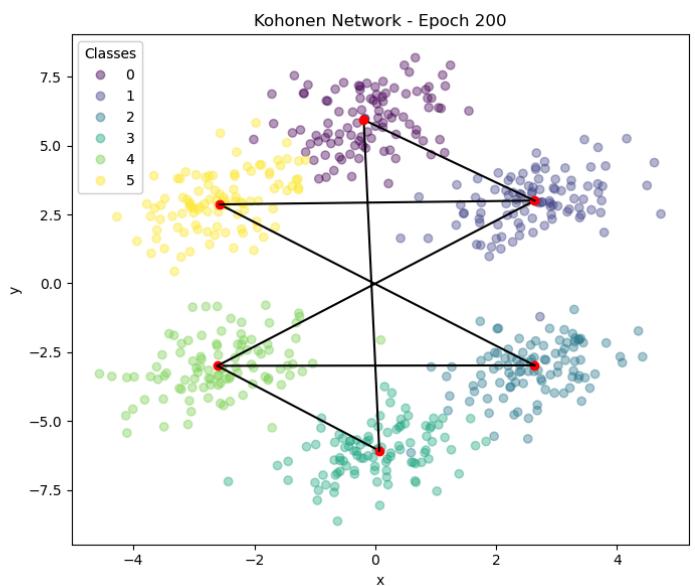
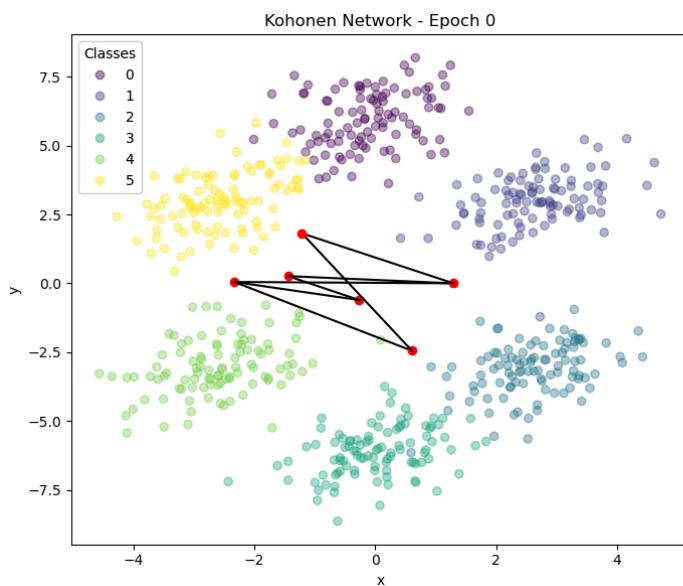


Rysunek 6: Najbliższe neuronów punkty, kolor: prawdziwa etykietka klasy.

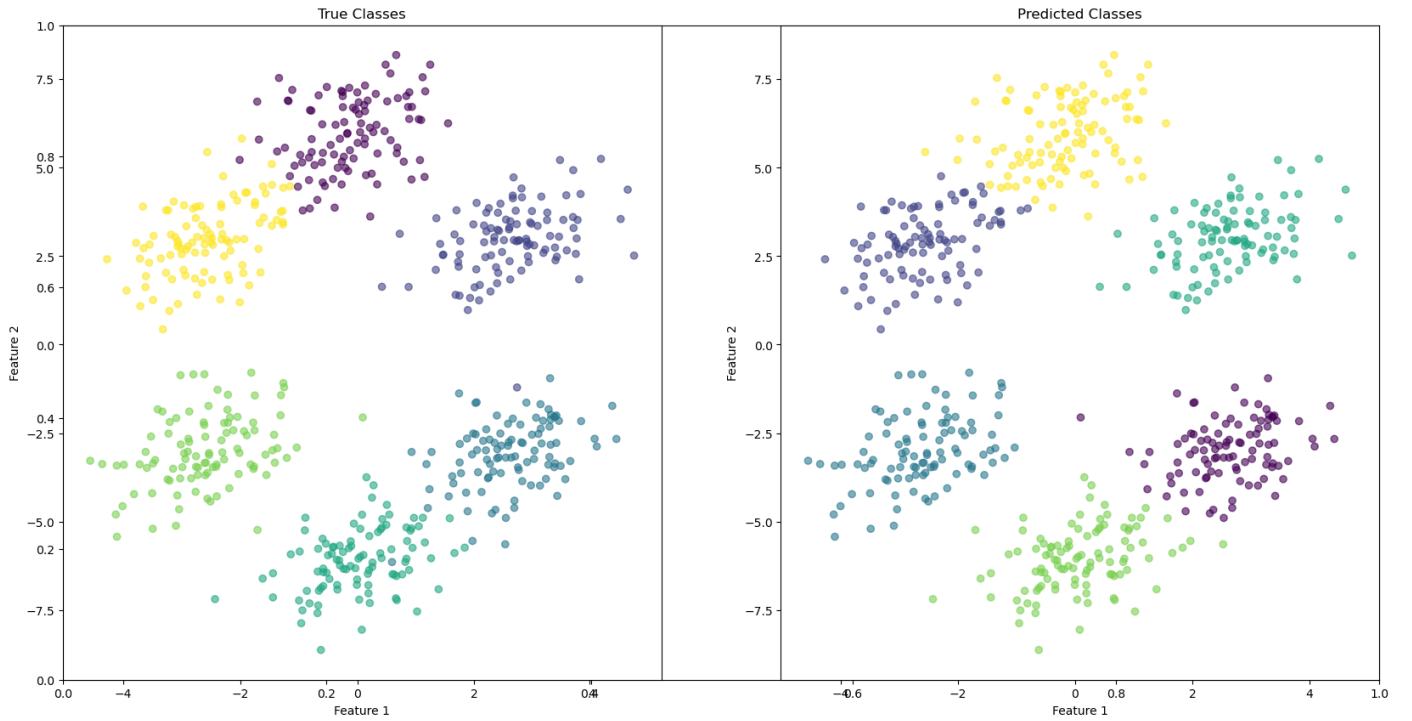


Rysunek 7: Najczęściej występująca prawdziwa klasa w neuronie.

według euklidesowskiej metryki. Ze względu na różnicę w etykietowaniu (wynikającą z numeracji neuronów), niemożliwe jest liczenie accuracy czy innych metryk dla zadań klasyfikacji. Istotnie, klasa oryginalnie fioletowa u nas jest żółta.



Rysunek 8: Progres uczenia się sieci. Czerwone punkty reprezentują neurony.



Rysunek 9: Klasyfikacja na podstawie (1) prawdziwych etykiet i (2) uczenia nienadzorowanego sieci Kohonena.

#### 4.1.2 Eksperymenty i wyniki

W prawdziwym zadaniu klasteryzacji z nieznaną liczbą klas należy zbadać sieć o innej liczbie neuronów. Posłużyono się rekomendowanym wzorem:

$$M \cdot N = 5\sqrt{n},$$

gdzie  $M \cdot N$  to liczba neuronów, a  $n$  to liczba próbek w zbiorze danych. W Hexagon jest 600 rekordów, więc wymiary siatki powinny wynosić  $11 \times 11$ .

Dla zbioru danych Hexagon zbadano oba rodzaje siatek, obie funkcje odległości, współczynniki sąsiedztwa 0.1, 0.5 i 1 oraz szybkość uczenia 0.001, 0.01, 0.1 i 1. Architektury oceniono względem indeksu Silhouette. Najlepszymi trzema są:

- Sieć 1: kwadratowa, szybkość uczenia: 1.0, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0;
- Sieć 2: kwadratowa, szybkość uczenia: 0.1, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0;
- Sieć 3: sześciokątna, szybkość uczenia: 1.0, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0.

Tabela 1: Wynik Silhouette a konfiguracja sieci dla danych Hexagon

	Gaussian Distance			Mexican Hat Distance		
	NS=0.1	NS=0.5	NS=1.0	NS=0.1	NS=0.5	NS=1.0
Sq & LR=0.001	0.177	0.196	0.176	0.231	0.193	0.192
Sq & LR=0.01	0.237	0.204	0.227	0.220	0.234	0.255
Sq & LR=0.1	0.294	0.293	0.262	0.322	0.327	0.546
Sq & LR=1.0	0.364	0.352	0.317	0.362	0.303	0.595
Hex & LR=0.001	0.193	0.213	0.221	0.212	0.194	0.169
Hex & LR=0.01	0.179	0.231	0.189	0.211	0.224	0.247
Hex & LR=0.1	0.297	0.295	0.265	0.299	0.321	0.341
Hex & LR=1.0	0.368	0.356	0.372	0.348	0.326	0.374

Trzy najlepsze sieci (zaznaczone kolorem w Tabeli 1) zostały poddane ewaluacji. Wyniki zebrane w Tabeli 2. Można wywnioskować, że hierarchia architektur nie zmieniła się po dodatkowych testach.

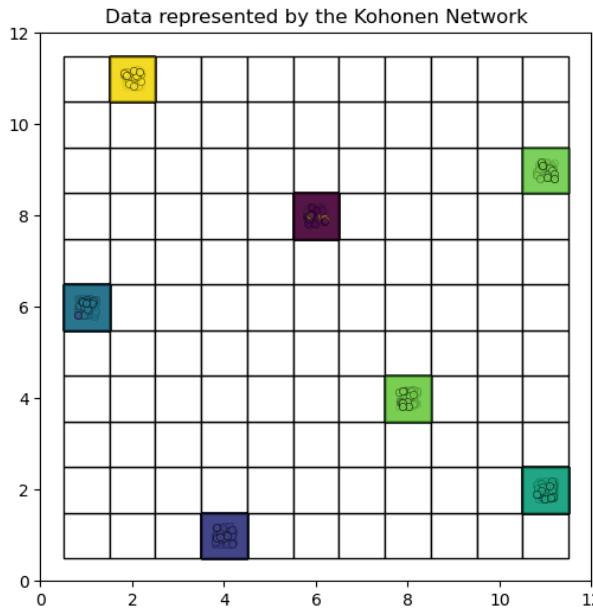
Tabela 2: Porównanie najlepszych sieci dla danych Hexagon.

Metryka	Sieć 1	Sieć 2	Sieć 3
Silhouette (Średnia)	0.551	0.540	0.393
Silhouette (Odch. Standardowe)	0.024	0.023	0.009
Davies-Bouldin (Średnia)	0.650	0.687	0.904
Davies-Bouldin (Odch. Standardowe)	0.070	0.056	0.015

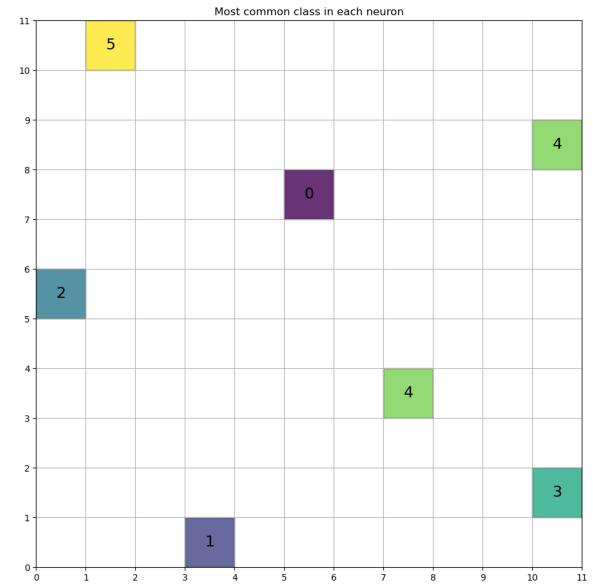
#### 4.1.3 Wizualizacje najlepszej sieci

Przedstawiona na Rysunkach 10, 11, 12 i 13 sieć Kohonena osiąga wyniki:

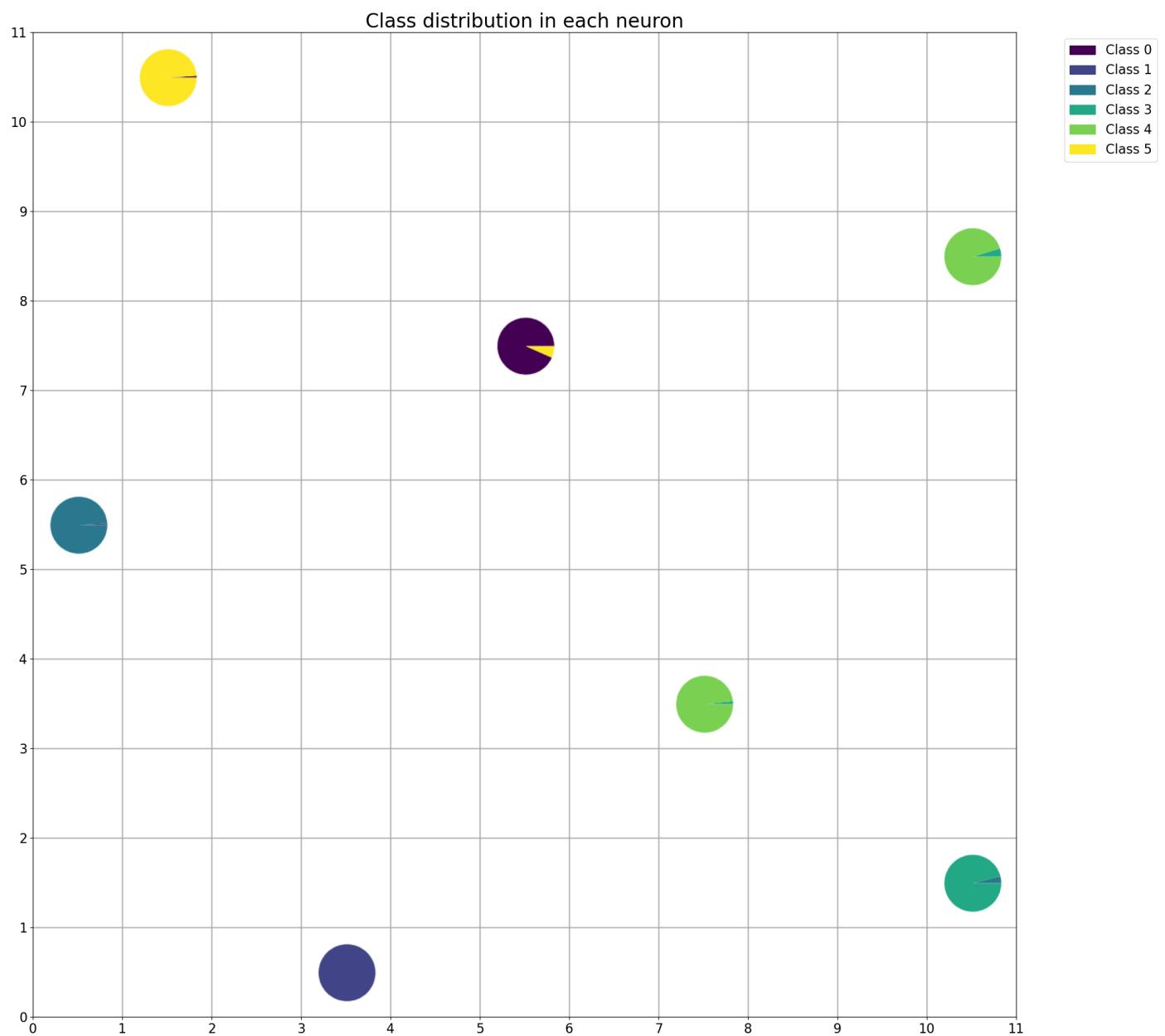
- Silhouette = 0.5544,
- Davies-Bouldin = 0.6738.



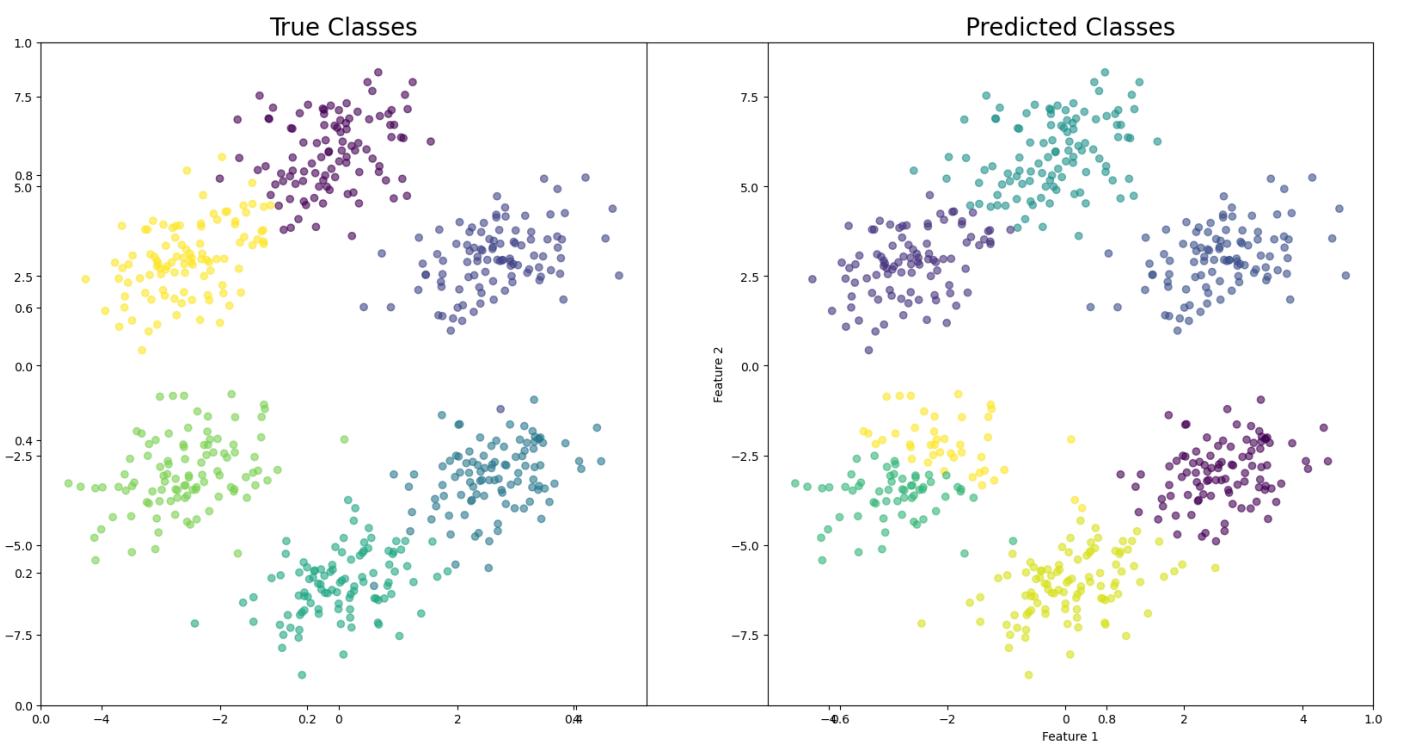
Rysunek 10: Najbliższe neuronów punkty, kolor: prawdziwa etykieta klasy. Dane: Hexagon.



Rysunek 11: Najczęściej występująca (prawdziwa) klasa w nauronie. Dane: Hexagon.



Rysunek 12: Udział (prawdziwych) klas w każdym z neuronów. Dane: Hexagon.



Rysunek 13: Klasyfikacja na podstawie (1) prawdziwych etykiet i (2) uczenia nienadzorowanego sieci Kohonena. Dane: Hexagon.

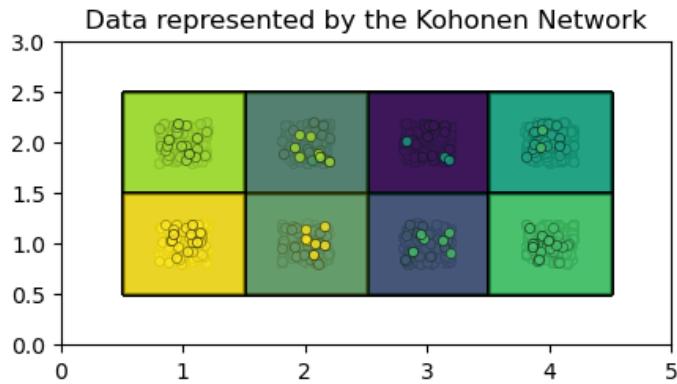
#### 4.1.4 Wnioski

Dla zbioru danych Hexagon preferowaną funkcją odległości jest MexicanHat, a typem siatki - kwadratowa.

## 4.2 Cube

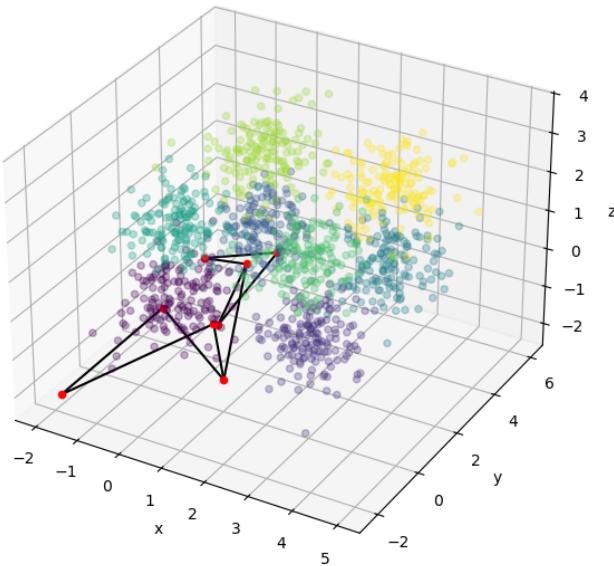
### 4.2.1 Przykładowa sieć

Weźmy sieć z 8-ma neuronami, o architekturze  $2 \times 4$ . Współczynnik sąsiedztwa został dobrany ręcznie. Wizualizacja klas punktów najbliższej neuronów znajduje się na Rysunku 14. Na Rysunku 15 przedstawiono postęp uczenia się sieci po 1000 epok.

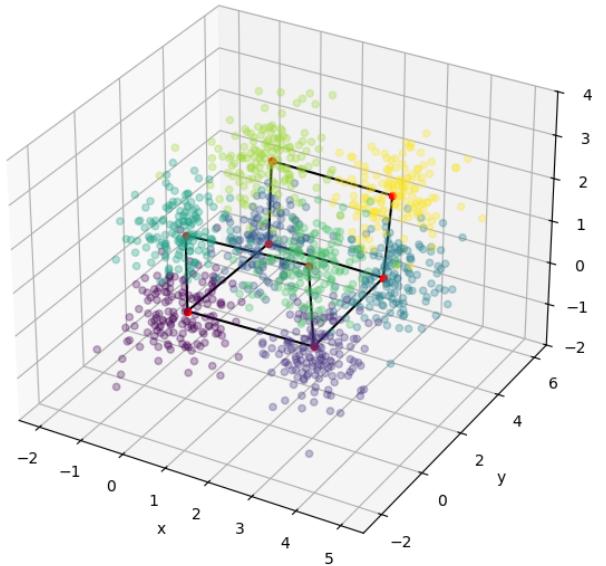


Rysunek 14: Najbliższe nauronów punkty, kolor: prawdziwa etykieta klasy.

Kohonen Network - Epoch 0



Kohonen Network - Epoch 1000



Rysunek 15: Postęp uczenia się sieci. Czerwone punkty reprezentują neurony.

Osiągnięte wyniki dla tej sieci: Silhouette = 0.4317 i Davis-Bouldin = 0.9103.

### 4.2.2 Eksperymenty i wyniki

Ponownie, rozważmy zadanie klasteryzacji, w którym nie znamy liczby klas. W tym celu, wykorzystując ten sam wzór, decydujemy się na siatkę  $13 \times 13$ . Pozostale badane parametry są takie, jak eksperymenty dla danych Hexagon. Wyniki doświadczenia zebrane w Tabeli 3.

Najlepszymi sieciami są:

- Sieć 1: kwadratowa, szybkość uczenia: 1.0, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0;
- Sieć 2: kwadratowa, szybkość uczenia: 0.1, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0;
- Sieć 3: sześciokątna, szybkość uczenia: 1.0, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0.

Tabela 3: Wynik Silhouette a konfiguracja sieci dla danych Cube

	Gaussian Distance			Mexican Hat Distance		
	NS=0.1	NS=0.5	NS=1.0	NS=0.1	NS=0.5	NS=1.0
<b>Sq &amp; LR=0.001</b>	0.135	0.125	0.119	0.109	0.127	0.088
<b>Sq &amp; LR=0.01</b>	0.130	0.150	0.073	0.142	0.137	0.230
<b>Sq &amp; LR=0.1</b>	0.221	0.198	0.168	0.240	0.235	0.355
<b>Sq &amp; LR=1.0</b>	0.252	0.247	0.217	0.248	0.220	0.416
<b>Hex &amp; LR=0.001</b>	0.122	0.140	0.105	0.121	0.139	0.111
<b>Hex &amp; LR=0.01</b>	0.165	0.149	0.102	0.149	0.167	0.160
<b>Hex &amp; LR=0.1</b>	0.228	0.226	0.186	0.237	0.215	0.244
<b>Hex &amp; LR=1.0</b>	0.253	0.257	0.252	0.246	0.204	0.263

Tabela 4: Porównanie najlepszych sieci dla danych Cube.

Metryka	Sieć 1	Sieć 2	Sieć 3
Średnia Silhouette	0.346	0.359	0.253
Odchylenie standardowe Silhouette	0.000	0.028	0.013
Średnia Davies-Bouldin	1.171	1.148	1.394
Odchylenie standardowe Davies-Bouldin	0.000	0.097	0.052

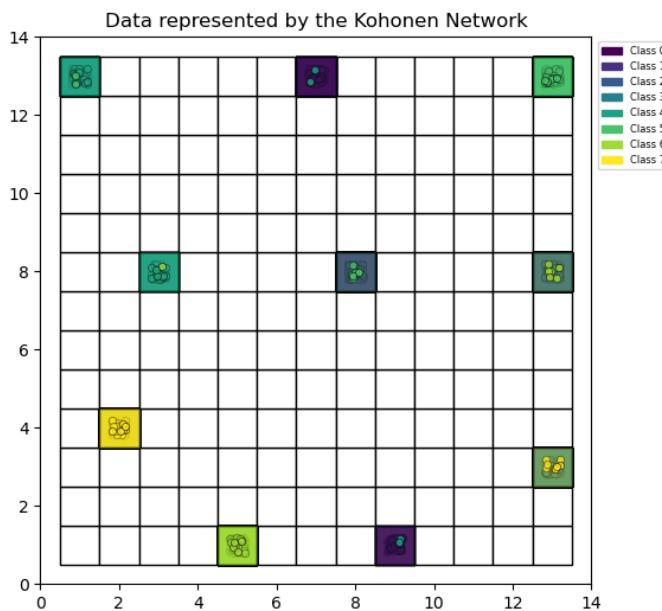
Wyniki dalszych testów znajdują się w Tabeli 4.

Gdy model przyporządkował wszystkie dane do jednego klastra, niemożliwe jest obliczenie metryk Silhouette i Davies-Bouldin. Takie sytuacje są pomijane w statystykach, z czego wynika zerowe odchylenie dla Sieci 1.

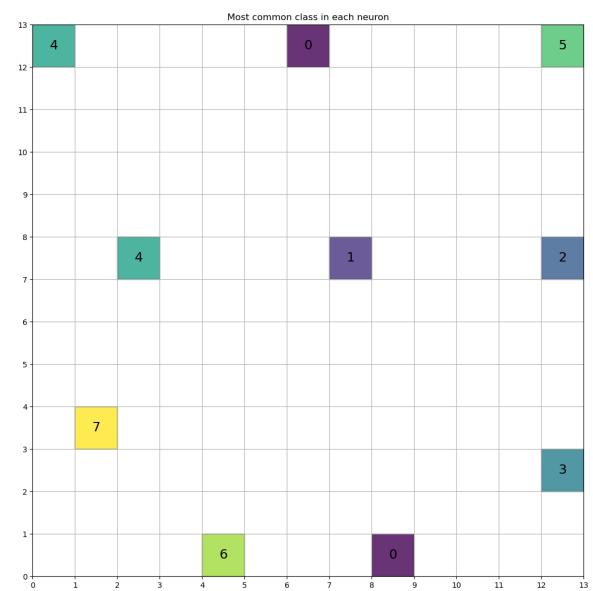
#### 4.2.3 Wizualizacje najlepszej sieci

Przedstawiona na Rysunkach 16, 17 i 18 sieć Kohonena osiąga wyniki:

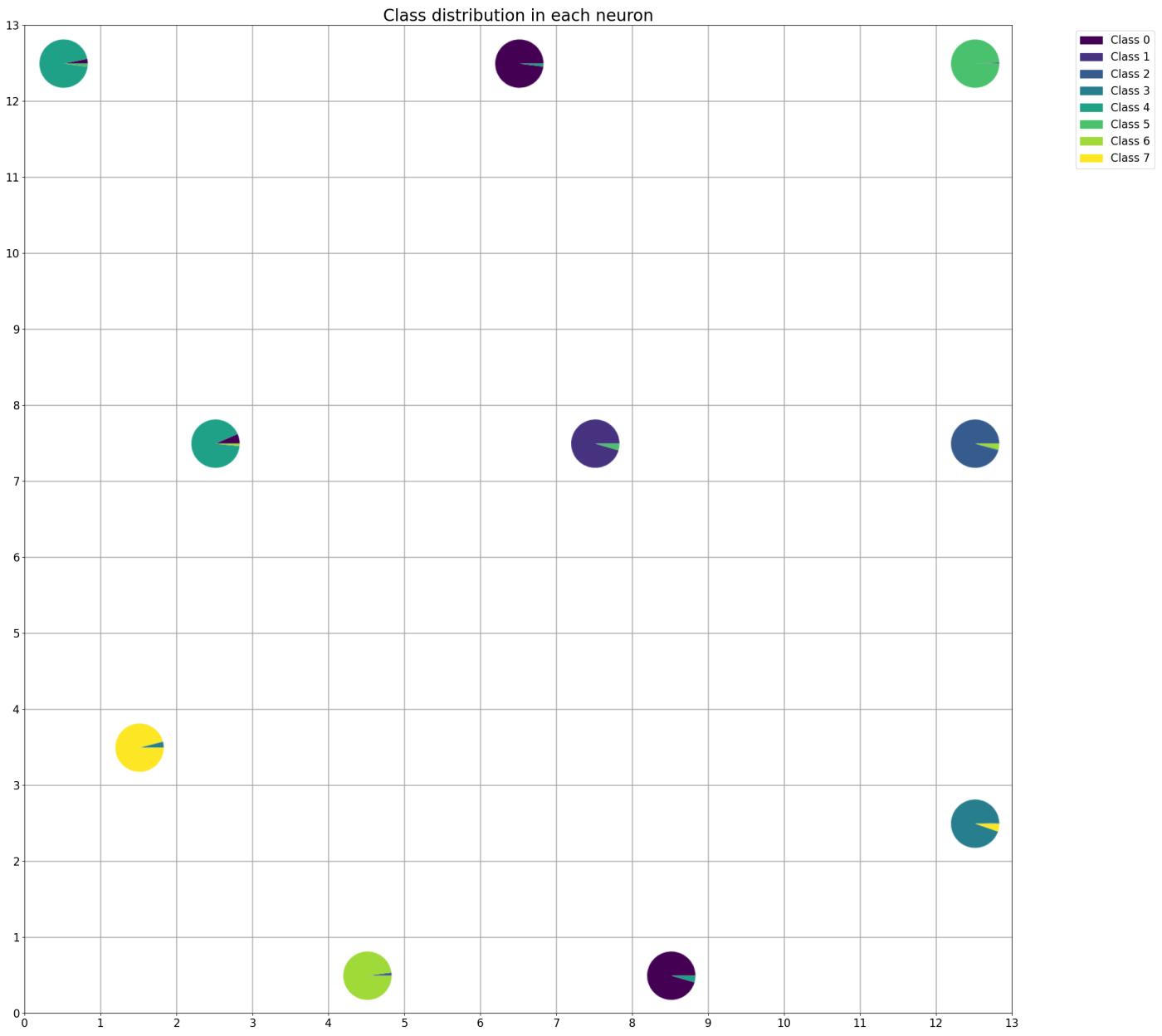
- Silhouette = 0.3833,
- Davies-Bouldin = 1.0829.



Rysunek 16: Najbliższe neuronów punkty, kolor: prawdziwa etykieta klasy. Dane: Cube.



Rysunek 17: Najczęściej występująca (prawdziwa) klasa w nauronie. Dane: Cube.



Rysunek 18: Udział (prawdziwych) klas w każdym z neuronów. Dane: Cube.

#### 4.2.4 Wnioski

Dla zbioru danych Cube, podobnie jak dla Hexagon, preferowaną funkcją odległości jest MexicanHat, a typem siatki - kwadratowa.

### 4.3 MNIST

#### 4.3.1 Wybór najlepszej sieci

Ponieważ zbiór MNIST ma 60 000 próbek, na potrzeby eksperymentów wykorzystamy pierwsze 2000. Zgodnie ze wzorem  $5\sqrt{n}$ , wybrano sieć  $14 \times 14$ . Wyniki zebrane w Tabeli 5. Wszystkie wartości są bliskie 0, co oznacza, że klastry nie są łatwo separowalne. Jest to spodziewany wynik, ponieważ wizualizacja t-SNE dla prawdziwych etykiet klas nie pozwalała na jednoznaczny podział grup punktów (Rysunek 4).

Najlepszymi sieciami w świetle metryki Silhouette są:

- Sieć 1: sześciokątna, szybkość uczenia: 1.0, funkcja odległości: Gaussian, współczynnik sąsiedztwa: 1.0;
- Sieć 2: kwadratowa, szybkość uczenia: 1.0, funkcja odległości: Gaussian, współczynnik sąsiedztwa: 0.1;
- Sieć 3: kwadratowa, szybkość uczenia: 1.0, funkcja odległości: Gaussian, współczynnik sąsiedztwa: 0.5.

Wyniki dalszych testów znajdują się w Tabeli 6.

Podobnie jak we wstępnych testach (Tabela 5), różnice między konfiguracjami nie są istotne.

Tabela 5: Wyniki Silhouette a konfiguracja dla danych MNIST.

	Gaussian Distance			Mexican Hat Distance		
	NS=0.1	NS=0.5	NS=1.0	NS=0.1	NS=0.5	NS=1.0
<b>Sq &amp; LR=0.001</b>	0.020	0.015	0.022	0.012	0.015	-0.019
<b>Sq &amp; LR=0.01</b>	0.003	0.011	0.027	0.014	-0.091	-0.145
<b>Sq &amp; LR=0.1</b>	0.031	0.061	0.050	0.032	0.023	0.061
<b>Sq &amp; LR=1.0</b>	<b>0.070</b>	<b>0.070</b>	0.057	0.064	0.064	0.067
<b>Hex &amp; LR=0.001</b>	0.019	0.016	0.017	0.008	0.011	0.001
<b>Hex &amp; LR=0.01</b>	0.018	0.013	0.027	0.015	-0.002	-0.105
<b>Hex &amp; LR=0.1</b>	0.045	0.040	0.060	0.034	0.014	0.060
<b>Hex &amp; LR=1.0</b>	0.067	0.066	<b>0.072</b>	0.069	0.066	0.059

Tabela 6: Porównanie najlepszych sieci dla danych MNIST.

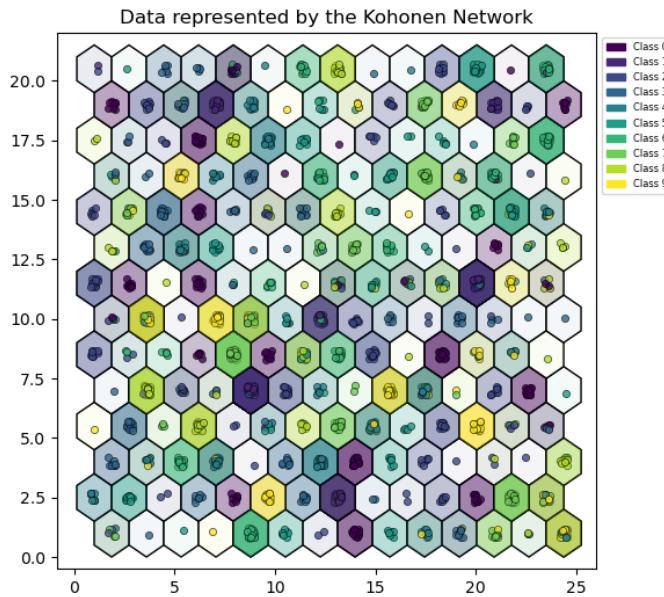
Metryka	Sieć 1	Sieć 2	Sieć 3
Średnia Silhouette	0.074	0.070	0.071
Odchylenie Silhouette	0.001	0.002	0.002
Średnia Davies-Bouldin	1.967	1.812	1.891
Odchylenie Davies-Bouldin	0.018	0.037	0.020

#### 4.3.2 Wizualizacje najlepszej sieci

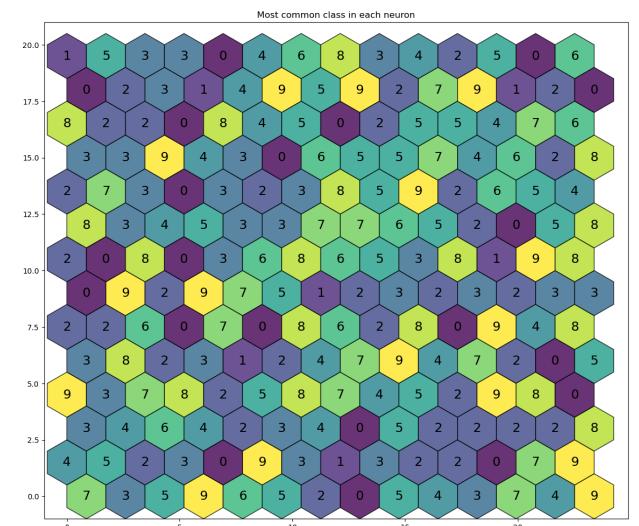
Przedstawiona na Rysunkach 19, 20, 21 i 22 sieć Kohonen osiąga wyniki:

- Silhouette = 0.0741,
- Davies-Bouldin = 1.8305.

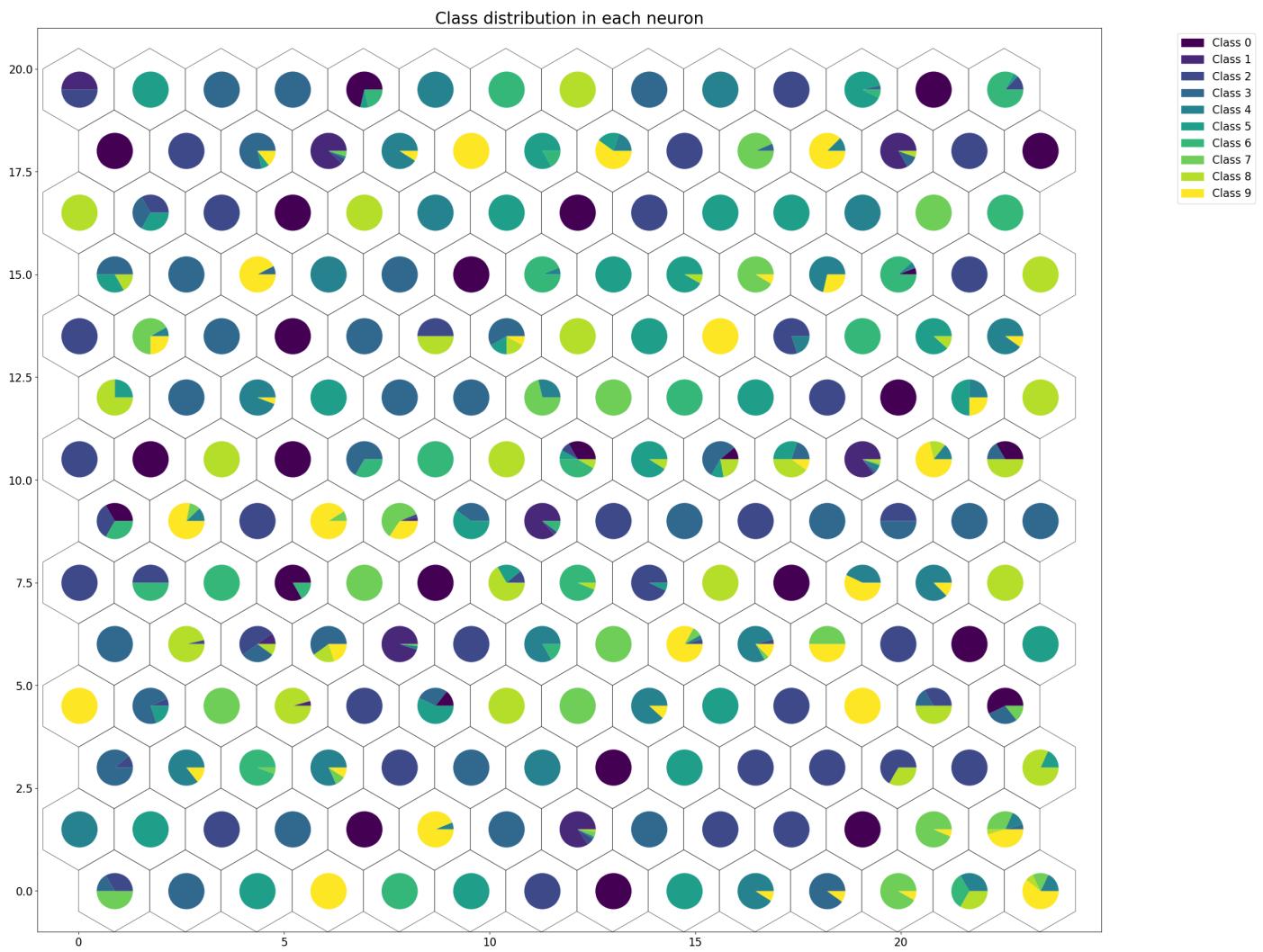
Uwaga. Na Rysunku 22 nie zamieszczono legendy z etykietami klas ze względu na ich dużą liczbę.



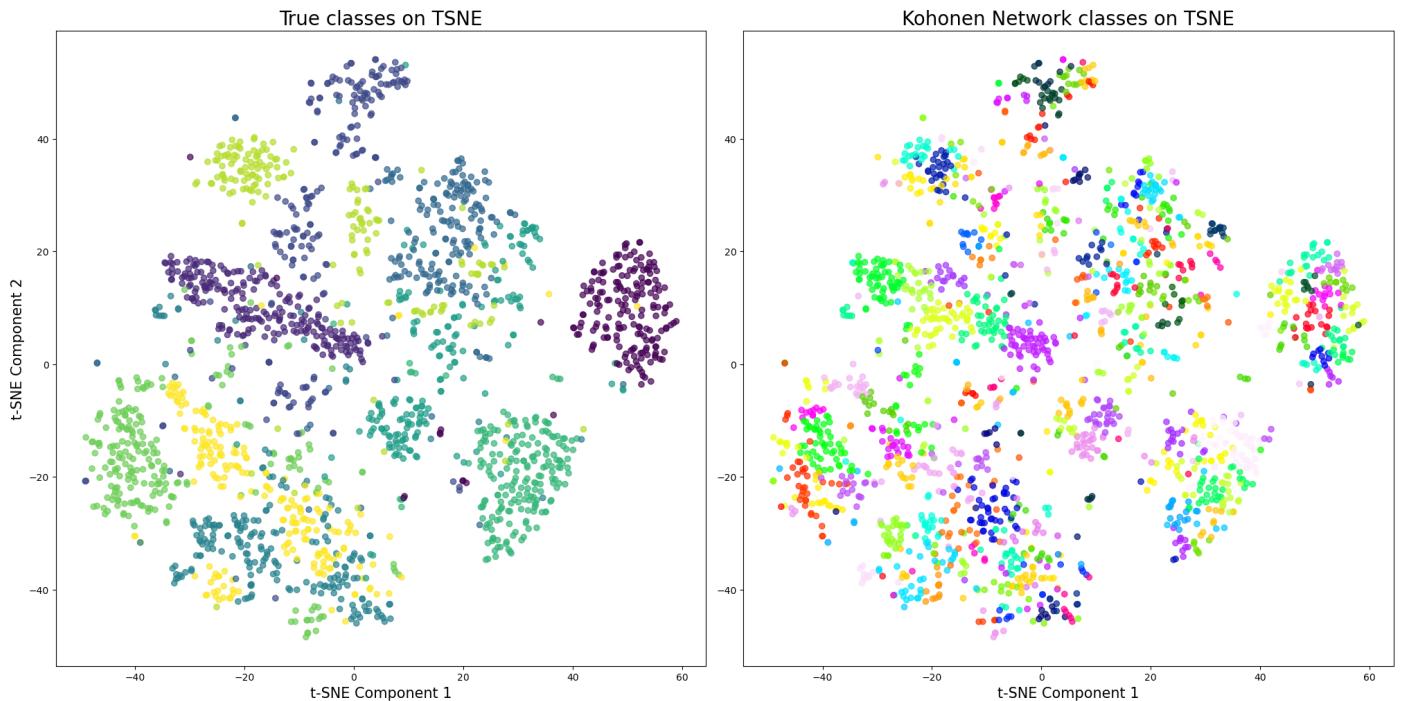
Rysunek 19: Najbliższe neuronów punkty, kolor: prawdziwa etykieta klasy. Dane: MNIST.



Rysunek 20: Najczęściej występująca (prawdziwa) klasa w nauronie. Dane: MNIST.



Rysunek 21: Udział (prawdziwych) klas w każdym z neuronów. Dane: MNIST.



Rysunek 22: Klasyfikacja na podstawie (1) prawdziwych etykiet i (2) uczenia nienadzorowanego sieci Kohonena. Wizualizacja względem dwóch komponentów t-SNE. Dane: MNIST.

### 4.3.3 Wnioski

Dla zbioru danych MNIST preferowaną funkcją odległości jest GaussianDistance. Dotychczas lepsze wyniki uzyskano wykorzystując MexicanHat. Wyniki klasteryzacji, zarówno metryki jak i wizualizacje, nie są w pełni zadowalające. Można się tego spodziewać po reprezentacji t-SNE kolorowanej prawdziwymi etykietami (Rysunek 4).

## 4.4 HARUS

### 4.4.1 Wybór najlepszej sieci

Zbiór HARUS zawiera ponad 7 000 próbek, dlatego na potrzeby doświadczenia wybrano pierwsze 2000. Stąd wymiary sieci są takie same jak dla danych MNIST.

Tabela 7: Wyniki Silhouette a konfiguracja dla danych HARUS.

	Gaussian Distance			Mexican Hat Distance		
	NS=0.1	NS=0.5	NS=1.0	NS=0.1	NS=0.5	NS=1.0
<b>Sq &amp; LR=0.001</b>	0.020	0.012	0.015	0.015	0.013	-0.061
<b>Sq &amp; LR=0.01</b>	0.011	0.019	0.045	0.021	-0.040	0.019
<b>Sq &amp; LR=0.1</b>	0.039	0.080	0.054	0.038	0.078	0.128
<b>Sq &amp; LR=1.0</b>	0.085	0.083	0.063	0.081	0.096	0.128
<b>Hex &amp; LR=0.001</b>	0.021	0.015	0.013	0.016	0.024	-0.011
<b>Hex &amp; LR=0.01</b>	0.021	0.024	0.020	0.021	0.021	-0.032
<b>Hex &amp; LR=0.1</b>	0.037	0.045	0.075	0.040	0.021	0.091
<b>Hex &amp; LR=1.0</b>	0.083	0.077	0.080	0.079	0.102	0.089

Najlepszymi sieciami w świetle metryki Silhouette są:

- Sieć 1: kwadratowa, szybkość uczenia: 1.0, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0;
- Sieć 2: kwadratowa, szybkość uczenia: 0.1, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 1.0;
- Sieć 3: sześciokątna, szybkość uczenia: 1.0, funkcja odległości: MexicanHat, współczynnik sąsiedztwa: 0.5.

Wyniki dalszych testów znajdują się w Tabeli 5.

Tabela 8: Porównanie najlepszych sieci dla danych HARUS.

Metryka	Sieć 1	Sieć 2	Sieć 3
Średnia Silhouette	–	0.125	0.093
Odchylenie Silhouette	–	0.009	0.004
Średnia Davies-Bouldin	–	2.235	2.142
Odchylenie Davies-Bouldin	–	0.072	0.032

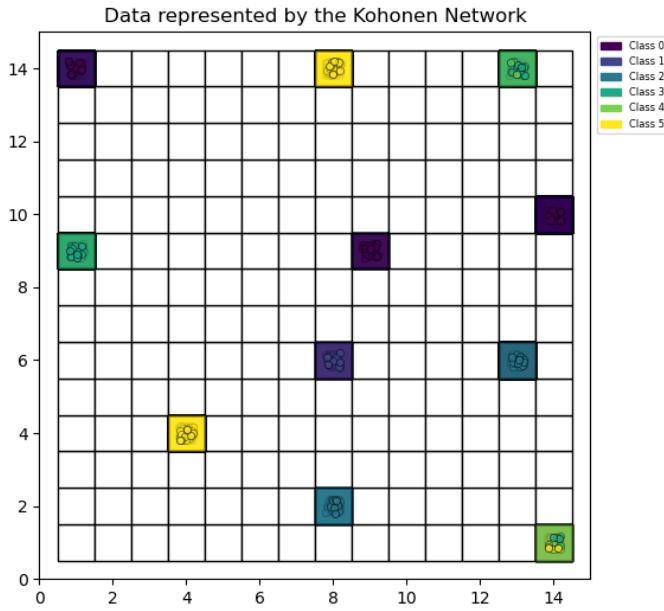
Dla sieci nr 1 nie mamy wyników. Jest to spowodowane faktem, że funkcja odległości MexicanHat jest wrażliwa na współczynnik *neighbourhood scaler*. Gdy dobierzemy zbyt dużą wartość tego parametru, może się zdarzyć, że dużo neuronów znacznie oddali się od obszaru z danymi. Wówczas często otrzymujemy jeden klaster. W naszym przypadku kombinacja współczynnika sąsiedztwa (1.0) i szybkości uczenia (1.0) doprowadziła do takiego zjawiska we wszystkich 5 iteracjach testów. Sieć nr 2 uzyskała stabilne wyniki we wszystkich testach.

### 4.4.2 Wizualizacje najlepszej sieci

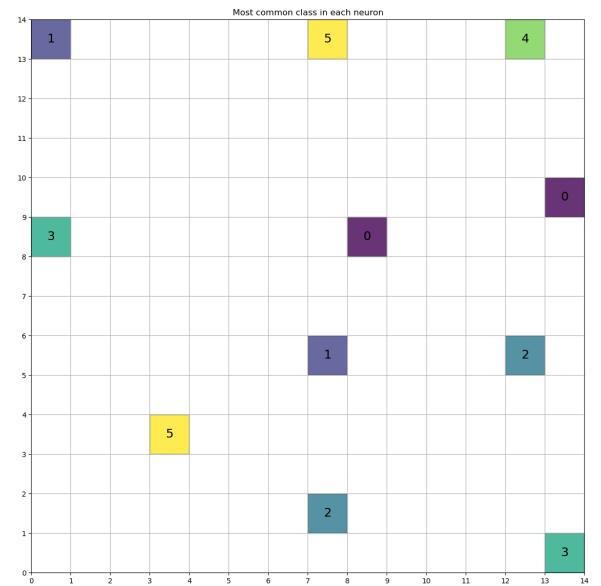
Przedstawiona na Rysunkach 23, 24, 25 i 26 sieć Kohonena osiąga wyniki:

- Silhouette = 0.1263,
- Davies-Bouldin = 2.2403.

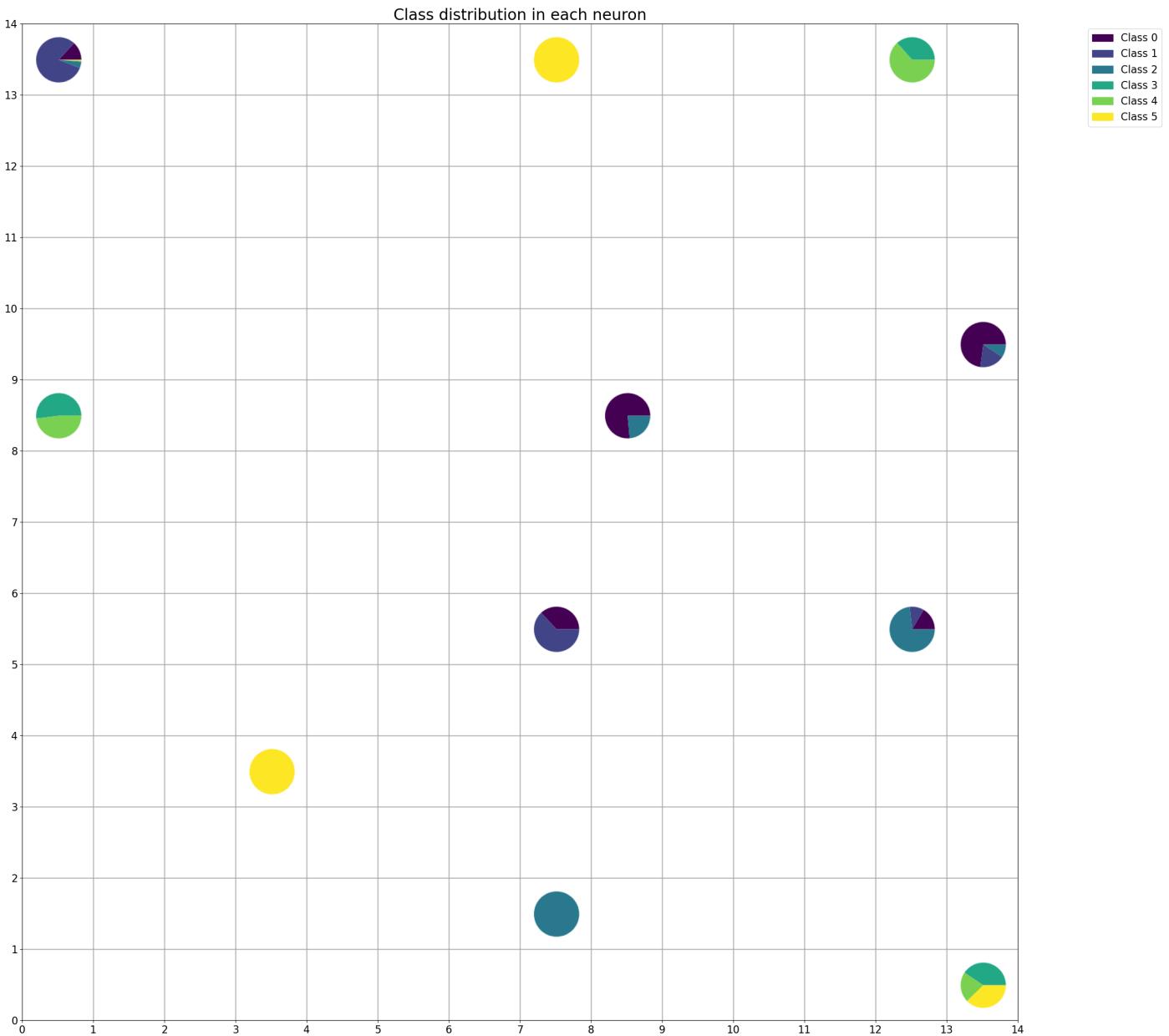
Uwaga. Na Rysunku 26 nie zamieszczono legendy z etykietami klas ze względu na ich dużą liczbę.



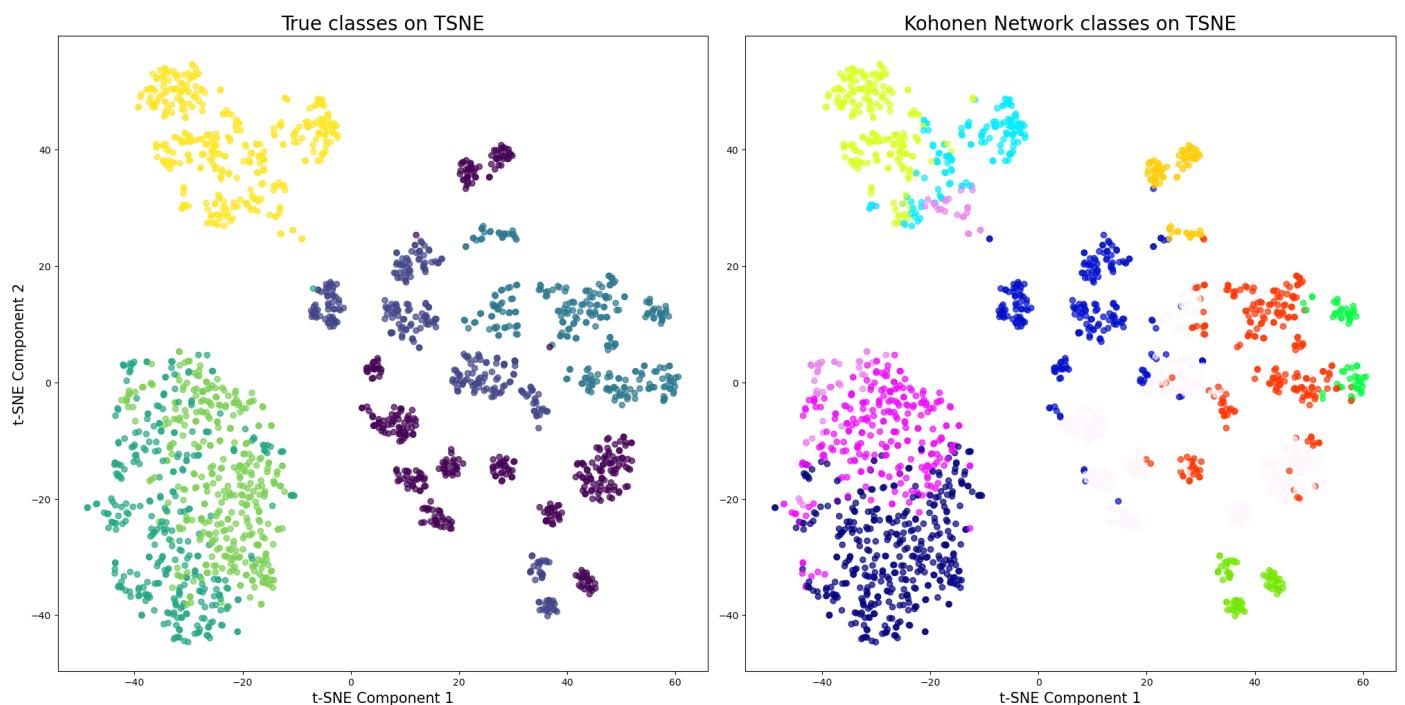
Rysunek 23: Najbliższe neuronów punkty, kolor: prawdziwa etykieta klasy. Dane: HARUS.



Rysunek 24: Najczęściej występująca (prawdziwa) klasa w neuronie. Dane: HARUS.



Rysunek 25: Udział (prawdziwych) klas w każdym z neuronów. Dane: HARUS.



Rysunek 26: Klasyfikacja na podstawie (1) prawdziwych etykiet i (2) uczenia nienadzorowanego sieci Kohonena. Wizualizacja względem dwóch komponentów t-SNE. Dane: HARUS.

#### **4.4.3 Wnioski**

Przy wyborze funkcji odległości MexicanHat bardzo istotny jest dobór właściwych wartości współczynników (sąsiedztwa i szybkości uczenia).

## **5 Podsumowanie**

Projekt pomógł stworzyć narzędzie służące do rozpoznawania klas oraz klasteryzacji. Przetestowano je na różnych zestawach danych i porównano wyniki. Ostatecznie można stwierdzić, że funkcja Mexican Hat lepiej spełniła oczekiwania - często osiągała lepsze wyniki. Jednakże istotnie trudniej jest dostosować jej parametry, aby osiągnąć zadowalające rezultaty. W niektórych przypadkach odległość Gaussowska uzyskiwała takie same lub lepsze metryki niż MexicanHat.

Z przeprowadzonych eksperymentów wynika, że nie ma uniwersalnego zestawu konfiguracyjnego, który działałby na każdym zbiorze danych.