# Zastosowanie uczenia nienadzorowanego do klasyfikacji

Filip Gaj

10 czerwca 2026

# Plan prezentacji

# Charakterystyka danych: Leukemia ALL/AML

W analizie wykorzystano klasyczny zbiór danych dotyczący białaczki (Golub et al.):

- **Cel:** Klasyfikacja typów białaczki: ALL (Ostra białaczka limfoblastyczna) oraz AML (Ostra białaczka szpikowa).
- **Liczba cech:** 7129 genów (bardzo wysoka wymiarowość).
- **Próbki:** Zbiór treningowy ($n = 38$) oraz testowy ($n = 34$).

# Wyzwania analizy ekspresji genów

Dane o ekspresji genów mają specyficzną i trudną strukturę:

**Klątwa wymiarowości ($p \gg n$)**

Zazwyczaj dysponujemy ogromną liczbą cech (genów, $p > 7\,000$), ale stosunkowo niewielką liczbą próbek (pacjentów, $n < 40$).

- Wysoki poziom szumu.
- Silna współliniowość.

# Preprocessing: Skalowanie danych

Geny mają różne poziomy ekspresji. Metody oparte na dystansie (K-Means) wymagają, aby wszystkie cechy miały tę samą wagę.

## 1. Standaryzacja (StandardScaler)

Przekształcenie danych do średniej $\mu = 0$ i odchylenia $\sigma = 1$:

$$z = \frac{x - \mu}{\sigma}$$

# Preprocessing: Skalowanie danych

## 2. Normalizacja (MinMaxScaler)

Skalowanie do przedziału $[0, 1]$:

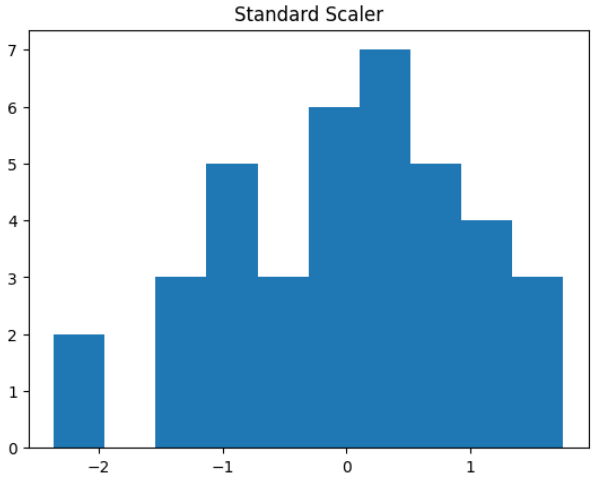
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

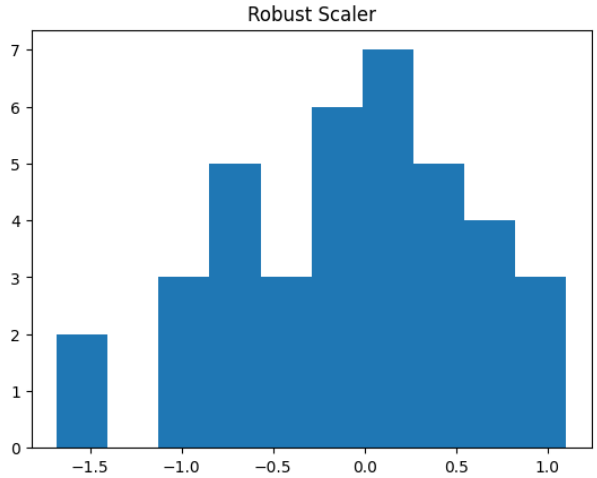
## 3. RobustScaler

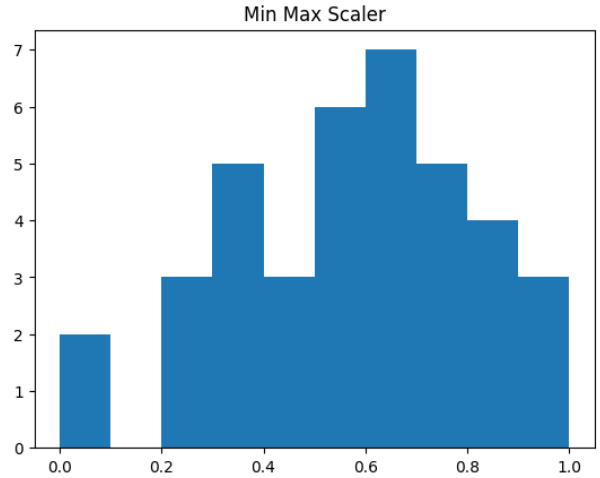
Używa mediany i rozstępu międzykwartylowego ($IQR = Q_3 - Q_1$):

$$x_{scaled} = \frac{x - \text{mediana}}{IQR}$$

# Porównanie metod skalowania danych



# Redukcja wymiarowości: PCA

Analiza Głównych Składowych (PCA) pozwala zredukować liczbę genów z 7129 do kilku najważniejszych składowych, zachowując maksymalną wariancję.

## Krok 1: Macierz kowariancji $C$

$$C = \frac{1}{n-1} X^T X$$

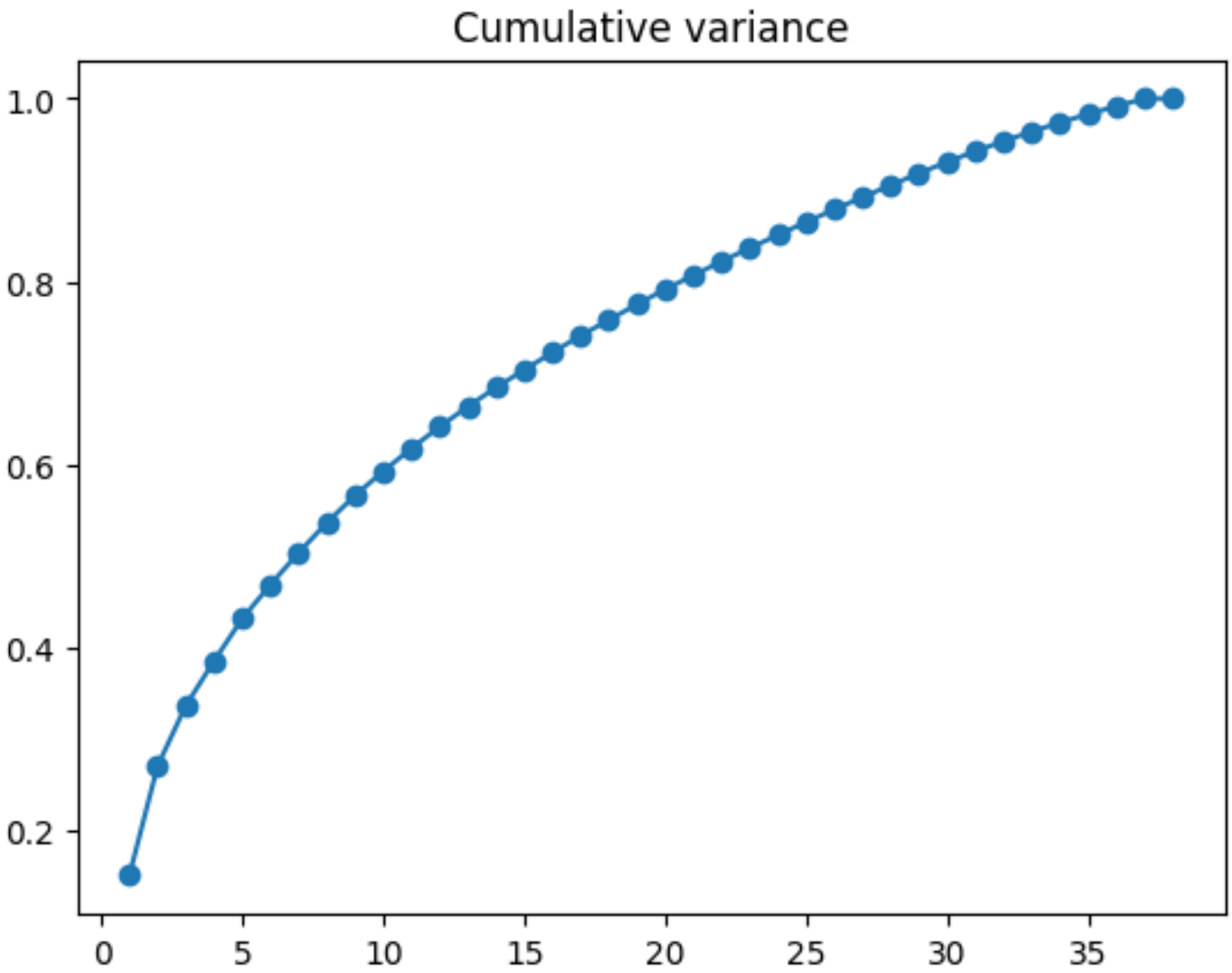
## Krok 2: Zagadnienie własne

Szukamy wektorów własnych $v$ i wartości własnych $\lambda$:

$$Cv = \lambda v$$

- **Główne składowe:** Wektory odpowiadające największym wartościom $\lambda$.
- **Wizualizacja:** Rzutowanie danych na płaszczyznę 2D utworzoną przez dwie pierwsze składowe.

# Wyjaśniana wariancja przez PCA



# t-SNE: t-distributed Stochastic Neighbor Embedding

Algorytm t-SNE przekształca podobieństwa między punktami na prawdopodobieństwa.

## 1. Podobieństwo w przestrzeni wysokowymiarowej ($p_{ij}$)

Prawdopodobieństwo, że punkt $x_j$ jest sąsiadem $x_i$ (rozkład Gaussa):

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}, \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

## 2. Podobieństwo w przestrzeni niskowymiarowej ($q_{ij}$)

Używamy rozkładu t-Studenta (1 stopień swobody), aby uniknąć crowding problem:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

# t-SNE: Funkcja kosztu

Celem t-SNE jest znalezienie takiego odwzorowania $y_i$, aby rozkład $Q$ był jak najbliższy rozkładowi $P$.

## Dywergencja Kullbacka-Leiblera (KL)

Minimalizujemy różnicę między rozkładami za pomocą gradientu:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- **Zaleta:** Świetna wizualizacja lokalnych skupisk (klastrów).
- **Wada:** Słabo zachowuje globalne relacje między odległymi klastrami i jest kosztowna obliczeniowo.

# UMAP: Uniform Manifold Approximation and Projection

Metoda UMAP działa ona w podobny sposób jak t-SNE, przy czym główną różnicą jest uwzględnienie składowej odpychającej, która ma na celu oddalenie od siebie punktów położonych daleko od siebie.

## 1. Lokalne podobieństwo w UMAP

Wykładniczy rozkład prawdopodobieństwa z uwzględnieniem odległości do najbliższego sąsiada ($\rho_i$):

$$w(x_i, x_j) = \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right)$$

## 2. Funkcja przynależności w niskim wymiarze

Podobnie jak w t-SNE, ale z parametrami $a$ i $b$ kontrolującymi zwartość klastrów:

$$\psi(y_i, y_j) = \left(1 + a(y_i - y_j)^{2b}\right)^{-1}$$

# UMAP: Optymalizacja

W przeciwieństwie do t-SNE, UMAP optymalizuje strukturę używając Cross-Entropy, co pozwala zachować zarówno lokalną, jak i globalną strukturę danych.

## Cross-Entropy (CE)

$$CE(P,Q) = \sum_{i \neq j} \left[ p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + (1-p_{ij}) \log\left(\frac{1-p_{ij}}{1-q_{ij}}\right) \right]$$

- **Część lewa:** Przyciąga punkty, które są blisko siebie (struktura lokalna).
- **Część prawa:** Odpycha punkty, które są daleko od siebie (struktura globalna).

# Klastrowanie (K-Means)

Klasyczny algorytm K-średnich dzieli pacjentów na $K$ grup, minimalizując wariancję wewnątrz klastrów.

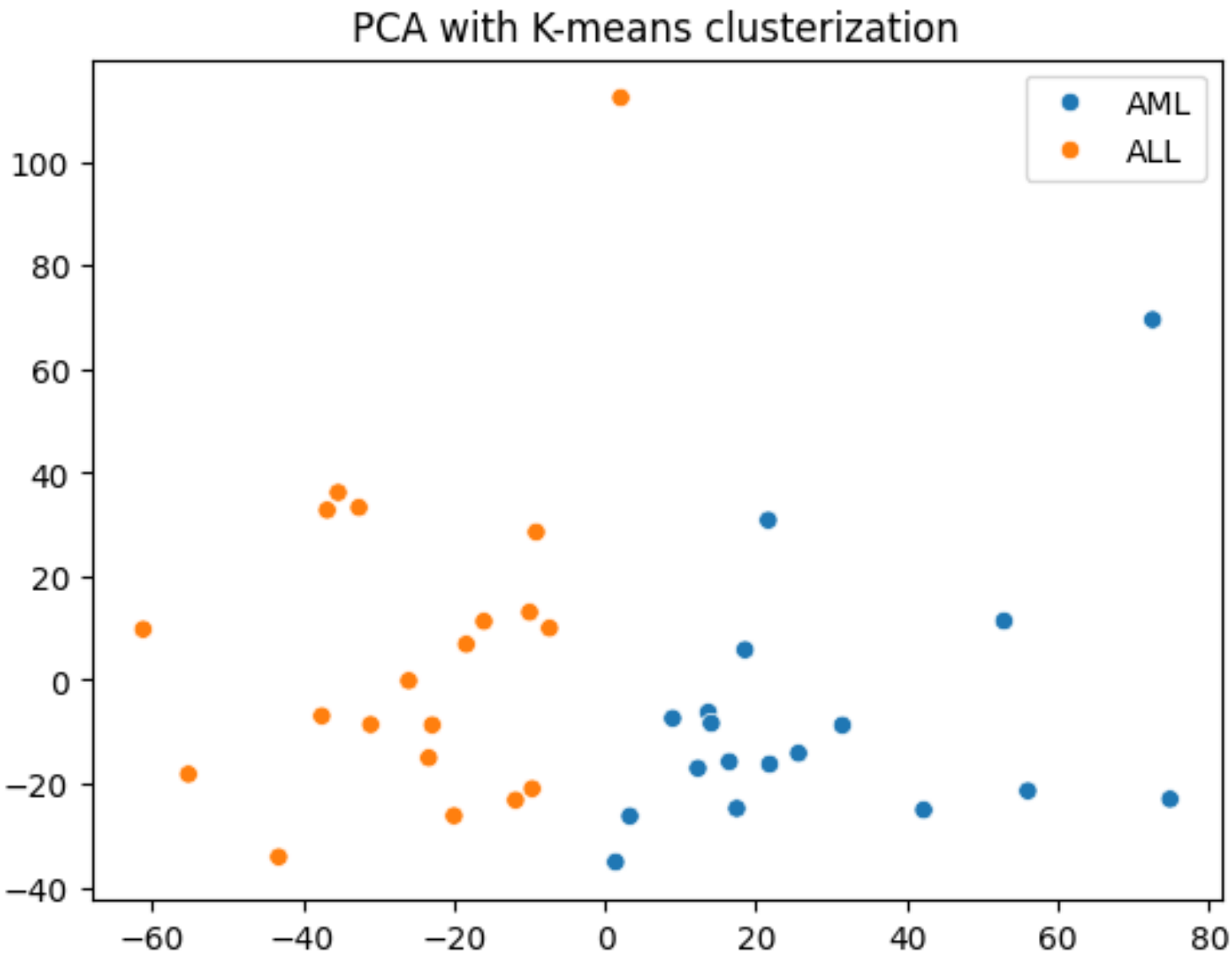
## Funkcja celu

Algorytm minimalizuje sumę kwadratów odległości euklidesowych:

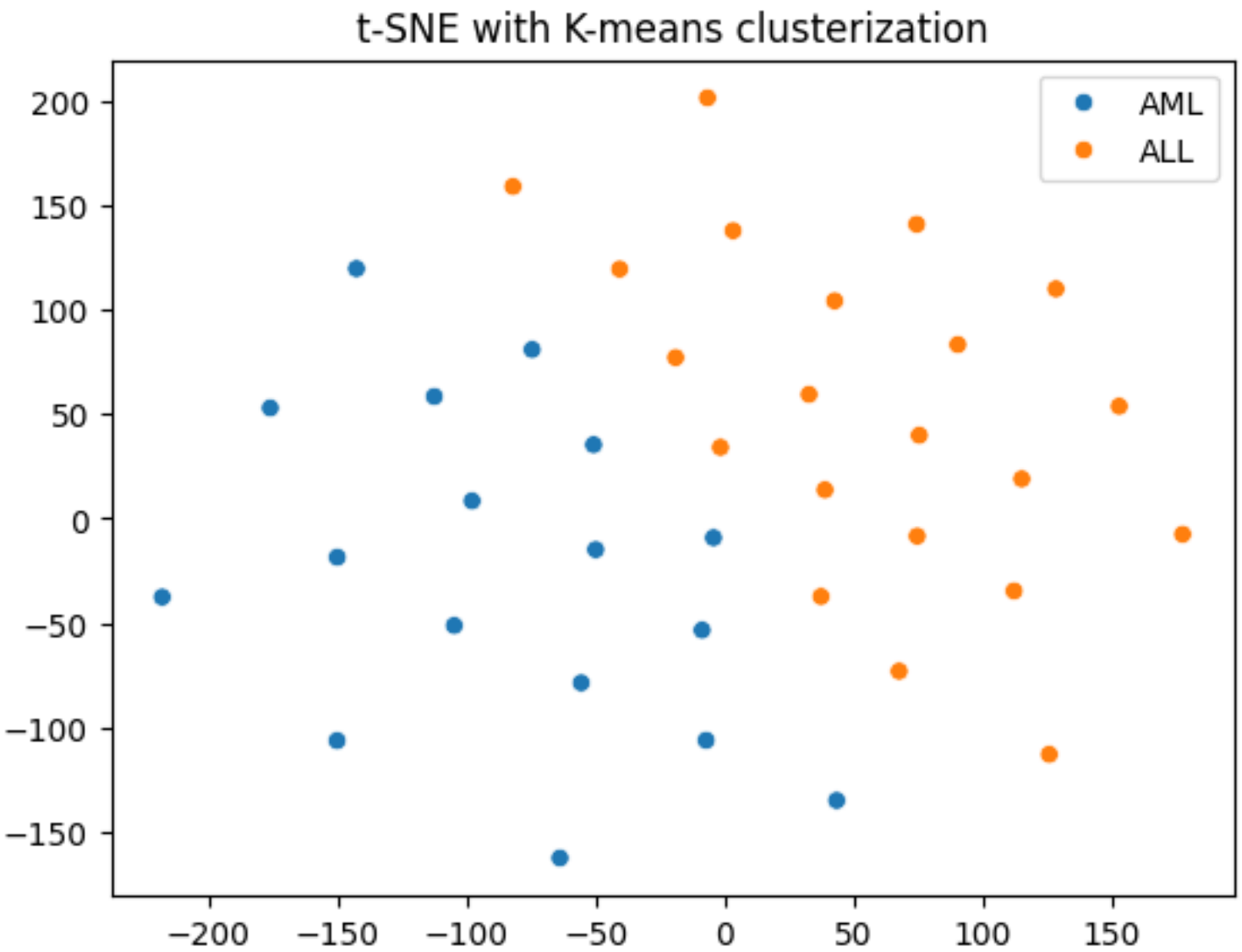
$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - \mu_j||^2 \tag{1}$$

gdzie $\mu_j$ to centroid (środek ciężkości) klastra $C_j$, a $x_i$ to wektor ekspresji genów pacjenta.

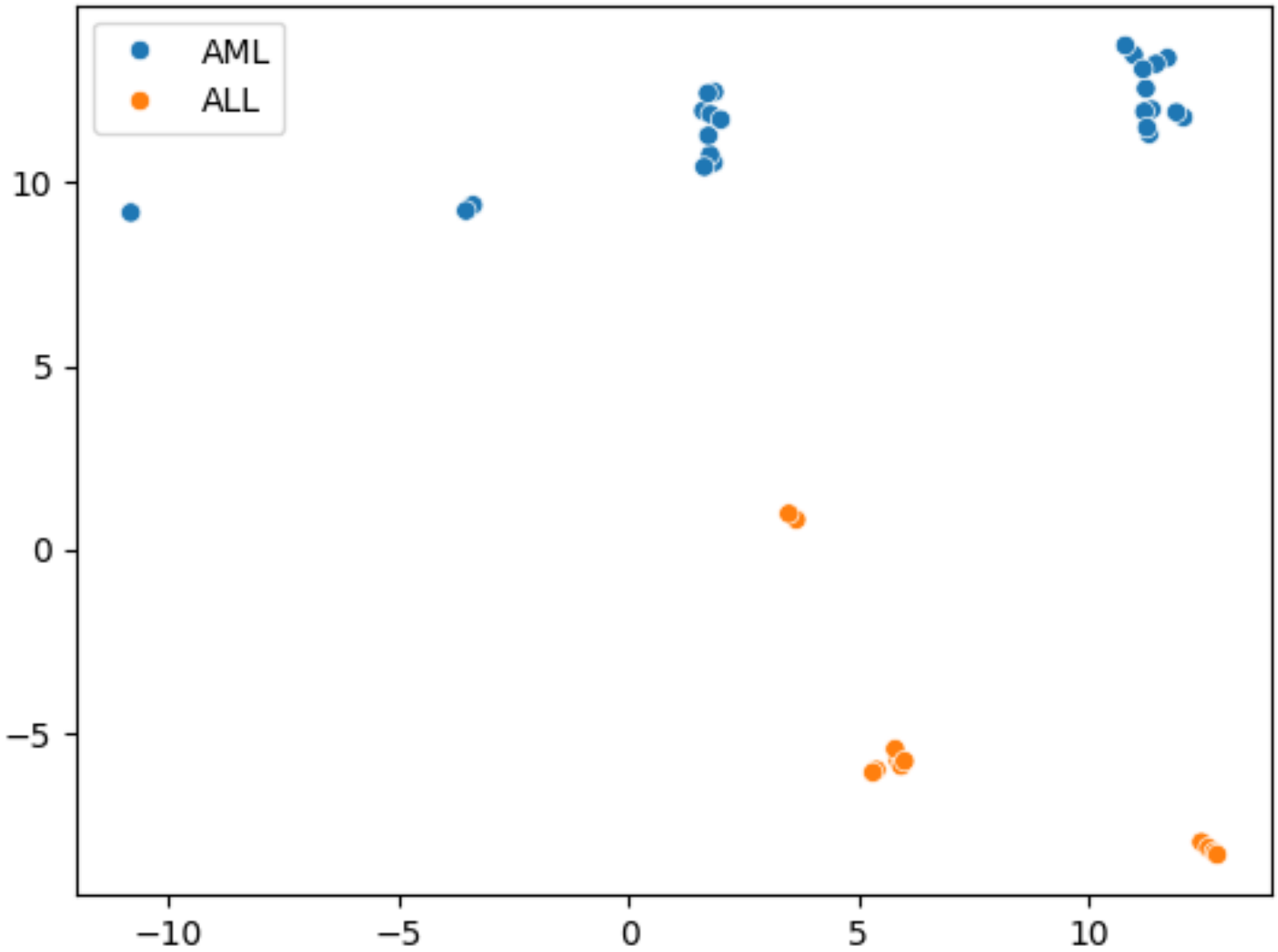
# Wizualizacja klastrów: PCA



# Wizualizacja klastrów: t-SNE



# Wizualizacja klastrów: UMAP



# Porównanie wyników klasyfikacji (PCA, t-SNE, UMAP)