Analiza Głównych Składowych (PCA) dla Zbioru Map Nacisku Stóp

Marcin Przybylski

2 maja 2025

Spis treści

1 Wstęp

Niniejszy dokument stanowi sprawozdanie z analizy zbioru danych map nacisku stóp przy użyciu metody Analizy Głównych Składowych (Principal Component Analysis - PCA). Celem ćwiczenia było poznanie i zrozumienie metody PCA, a następnie zastosowanie zdobytej wiedzy do analizy rzeczywistego zbioru danych biomedycznych. Analiza opierała się na danych wejściowych w postaci obrazów reprezentujących rozkład nacisku stopy na podłoże, pochodzących z pomiarów wykonanych dla różnych pozycji ciała (przysiad, skłon, stanie).

Główne cele analizy obejmowały:

- Zrozumienie struktury danych i podstawowych charakterystyk zbioru map nacisku stóp.
- Zastosowanie PCA do redukcji wymiarowości danych, przy jednoczesnym zachowaniu jak największej ilości informacji (wariancji).
- Wizualizację głównych składowych ("Eigenfeet") w celu zrozumienia głównych kierunków zmienności w danych.
- Ocenę jakości rekonstrukcji danych na podstawie zredukowanej reprezentacji.
- Identyfikację najbardziej charakterystycznych (najbardziej różniących się od siebie) map nacisku stóp w zbiorze za pomocą analizy odległości w przestrzeni PCA.

Prace zostały zrealizowane w środowisku Google Colaboratory, wykorzystując standardowe biblioteki języka Python do analizy danych i uczenia maszynowego, takie jak NumPy, Matplotlib oraz Scikit-learn.

Pełna instrukcja do ćwiczenia oraz notatnik Google Colab zawierający kod źródłowy, wyniki pośrednie i wizualizacje są dostępne pod poniższymi linkami:

- Instrukcja do ćwiczenia (Google Docs): Link do Instrukcji
- Roboczy notatnik Google Colab: Link do Colab

W niniejszym sprawozdaniu przedstawiono metodykę postępowania, uzyskane wyniki wraz z ich interpretacją oraz wnioski płynące z przeprowadzonej analizy PCA.

2 Podstawy Teoretyczne PCA

2.1 Czym jest Analiza Głównych Składowych (PCA)?

Analiza Głównych Składowych (PCA) to popularna technika statystyczna i algorytm uczenia maszynowego bez nadzoru, stosowany głównie do redukcji wymiarowości danych przy jednoczesnym zachowaniu jak największej ilości informacji (mierzonej jako wariancja) obecnej w oryginalnym zbiorze. Działa poprzez transformację liniową oryginalnych, potencjalnie skorelowanych zmiennych, w nowy zestaw nieskorelowanych zmiennych, zwanych głównymi składowymi. Składowe te są uporządkowane malejąco według ilości wariancji, którą wyjaśniają. Pierwsza główna składowa wyjaśnia największą część wariancji, druga składowa (ortogonalna do pierwszej) wyjaśnia największą część pozostałej wariancji itd.

Główne kroki algorytmu PCA to:

1. **Standaryzacja danych (opcjonalnie):** Przeskalowanie danych tak, aby każda cecha miała średnią równą 0 i odchylenie standardowe równe 1. Jest to ważne, gdy cechy mają różne skale. W analizowanym zbiorze stóp dane były już wstępnie przeskalowane.
2. **Obliczenie macierzy kowariancji (lub korelacji):** Określenie, jak zmienne współzmieniają się ze sobą. Macierz kowariancji jest obliczana dla danych standaryzowanych.
3. **Obliczenie wartości własnych i wektorów własnych:** Znalezienie wartości własnych (λ_i) i odpowiadających im wektorów własnych (v_i) macierzy kowariancji. Wektory własne wskazują kierunki maksymalnej wariancji w danych (są to osie głównych składowych), a wartości własne określają wielkość tej wariancji wzdłuż danego kierunku.
4. **Sortowanie głównych składowych:** Uporządkowanie wektorów własnych według malejących wartości własnych ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, gdzie p to liczba oryginalnych wymiarów).
5. **Wybór liczby składowych (k):** Decyzja, ile głównych składowych (o największych wartościach własnych) zachować ($k \leq p$). Wybór często opiera się na progu wyjaśnionej wariancji (np. zachowaj tyle składowych, aby wyjaśnić 95% całkowitej wariancji) lub analizie wykresu osypiska (scree plot).
6. **Transformacja danych:** Rzutowanie oryginalnych (standaryzowanych) danych na nową podprzestrzeń zdefiniowaną przez k wybranych wektorów własnych. Nowe współrzędne danych w tej podprzestrzeni to wartości głównych składowych.

2.2 Zalety PCA

- **Redukcja wymiarowości:** Skutecznie zmniejsza liczbę zmiennych, co może ułatwić wizualizację, przyspieszyć obliczenia w dalszych etapach analizy (np. w modelach uczenia maszynowego) i zredukować ryzyko przeuczenia.
- **Usuwanie korelacji:** Transformuje skorelowane zmienne w nieskorelowane składowe główne, co może być korzystne dla niektórych algorytmów.
- **Redukcja szumu:** Może pomóc w usunięciu szumu poprzez odrzucenie składowych o małej wariancji, które często reprezentują szum.
- **Identyfikacja wzorców:** Pomaga odkryć ukrytą strukturę i główne kierunki zmienności w danych.

2.3 Wady PCA

- **Wrażliwość na skalę:** Wyniki są zależne od skali zmiennych; standaryzacja danych jest często konieczna, aby uniknąć dominacji zmiennych o dużej wariancji.
- **Interpretowalność:** Główne składowe są kombinacjami liniowymi oryginalnych zmiennych, co może utrudniać ich bezpośrednią interpretację w kontekście problemu.
- **Założenie o liniowości:** PCA zakłada, że główne kierunki wariancji są liniowe i że interesująca struktura danych leży w kierunkach największej wariancji. Może nie

działać dobrze dla danych o złożonych, nieliniowych strukturach.

- **Utrata informacji:** Redukcja wymiarowości ($k < p$) zawsze wiąże się z pewną utratą informacji (wariancji). Należy starannie wybrać k, aby zminimalizować tę stratę.

2.4 Przykłady Zastosowań PCA

- **Rozpoznawanie twarzy (Eigenfaces):** Klasyczne zastosowanie PCA do redukcji wymiaru zdjęć twarzy i ekstrakcji cech charakterystycznych ("twarzy własnych"), które mogą być używane do identyfikacji lub klasyfikacji.
- **Analiza danych genetycznych:** Redukcja wymiarowości danych ekspresji genów (z tysięcy genów do kilku głównych składowych) w celu identyfikacji wzorców, grupowania próbek lub wizualizacji danych.
- **Przetwarzanie obrazów**: Kompresja obrazów poprzez zachowanie tylko najważniejszych składowych, co pozwala na zmniejszenie rozmiaru pliku przy akceptowalnej utracie jakości.
- **Finanse**: Analiza portfeli inwestycyjnych, identyfikacja głównych czynników ryzyka rynkowego wpływających na ceny aktywów.
- **Analiza danych sensorycznych**: Redukcja wymiarowości danych z wielu czujników w celu identyfikacji głównych trendów, anomalii lub stanu systemu.

3 Metodologia Analizy Danych Stóp

3.1 Wczytywanie i Przygotowanie Danych

Dane wejściowe, składające się z map nacisku stóp, zostały pobrane ze zdalnego serwera w formacie pliku `.npy` przy użyciu polecenia `wget`. Wczytano je do tablicy NumPy.

Dane miały oryginalny kształt `(18, 34, 6588)`, gdzie:

- 18: Wysokość obrazu mapy nacisku w pikselach.
- 34: Szerokość obrazu mapy nacisku w pikselach.
- 6588: Liczba dostępnych obrazów (próbek).

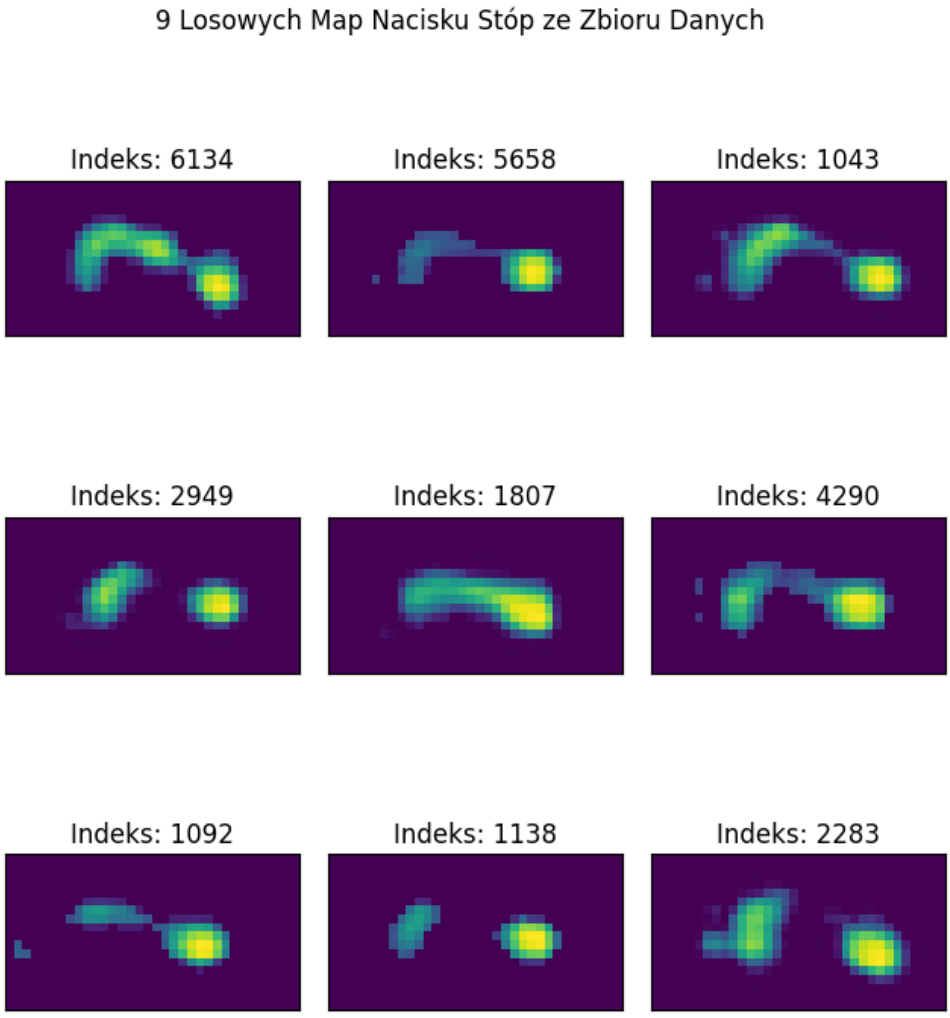
Do dalszej analizy konieczna była zmiana kolejności wymiarów, aby uzyskać format (liczba_próbek, wysokość, szerokość). Zastosowano funkcję `np.transpose(mapy_features, (2, 0, 1))`.

Po transpozycji uzyskano tablicę o kształcie `(6588, 18, 34)`. Podstawowe parametry zbioru danych:

- Liczba obrazów (stóp): $n_{samples} = 6588$.
- Wymiary obrazu (wysokość x szerokość): $h \times w = 18 \times 34$ pikseli.
- Liczba cech (pikseli) na obraz po spłaszczeniu: $n_{features} = h \times w = 612$.

Na potrzeby algorytmu PCA, który operuje na danych w formacie (liczba_próbek, liczba_cech), obrazy zostały spłaszczone za pomocą metody `.reshape()`. Ostateczny kształt danych wejściowych do PCA to `(6588, 612)`.

Przykładowe mapy nacisku stóp zostały zwizualizowane w celu wstępnej oceny danych (Rysunek 1).



Rysunek 1: Przykładowe mapy nacisku stóp ze zbioru danych.

3.2 Implementacja PCA

Analizę PCA przeprowadzono przy użyciu klasy `PCA` z modułu `sklearn.decomposition`. Kluczowe kroki implementacyjne:

1. Utworzenie instancji `PCA(n_components=None)`, aby najpierw obliczyć wszystkie możliwe składowe główne i zbadać wyjaśnioną wariancję.
2. Dopasowanie modelu do danych spłaszczonych: `pca_full.fit(X_stopy)`.
3. Analiza skumulowanej wyjaśnionej wariancji (`cumulative_explained_variance`) i wyznaczenie liczby składowych potrzebnych do wyjaśnienia określonego progu wariancji (np. 95%).

4. Utworzenie nowej instancji PCA z wybraną liczbą składowych:

   `PCA(n_components=n_components_selected)`.

5. Dopasowanie finalnego modelu PCA: `pca.fit(X_stopy)`.

6. Transformacja oryginalnych danych do przestrzeni o zredukowanej wymiarowości: `X_stopy_pca = pca.transform(X_stopy)`.

7. Odwrotna transformacja (rekonstrukcja) danych:

   `X_stopy_reconstructed = pca.inverse_transform(X_stopy_pca)`.

3.3 Identyfikacja Charakterystycznych Stóp

W celu znalezienia najbardziej różniących się od siebie map nacisku stóp, obliczono odległości euklidesowe między wszystkimi parami próbek w przestrzeni o zredukowanej wymiarowości (`X_stopy_pca`). Zgodnie z instrukcją, obliczenia przeprowadzono dwukrotnie, używając:

- `sklearn.metrics.pairwise.euclidean_distances`
- `scipy.spatial.distance.pdist` (z metryką 'euclidean') i

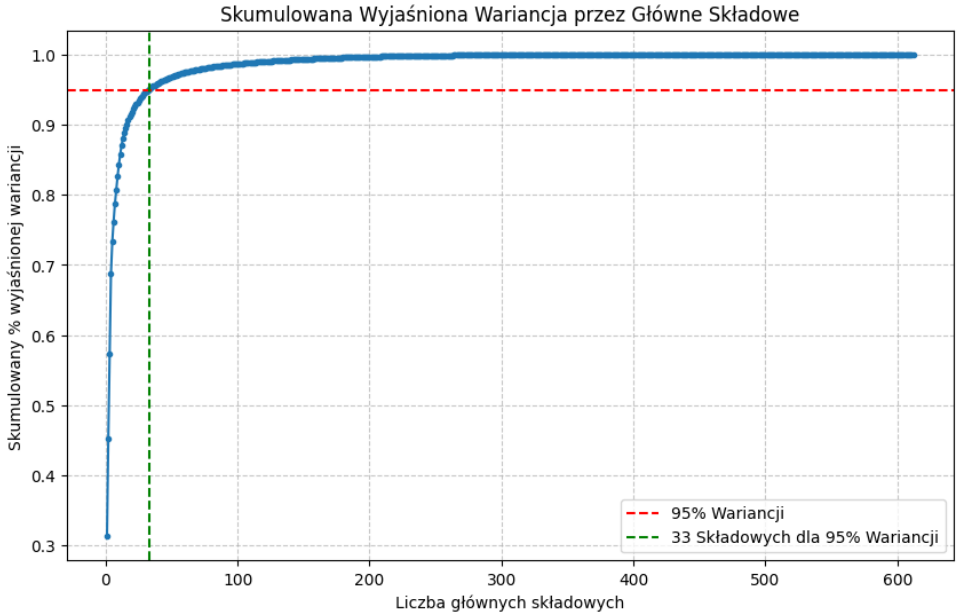
  `scipy.spatial.distance.squareform`

Następnie sprawdzono zgodność wyników obu metod i zidentyfikowano parę indeksów próbek (`idx1, idx2`), dla których odległość była największa.

4 Wyniki i Interpretacja

4.1 Analiza Wyjaśnionej Wariancji

Pierwszym krokiem analizy PCA było zbadanie, jak poszczególne składowe główne przyczyniają się do całkowitej wariancji w danych. Obliczono 612 składowych głównych (równo liczbie oryginalnych cech). Wykres skumulowanej wyjaśnionej wariancji (Rysunek 2) pokazuje, że znacząca część wariancji jest skoncentrowana w pierwszych kilkudziesięciu składowych.



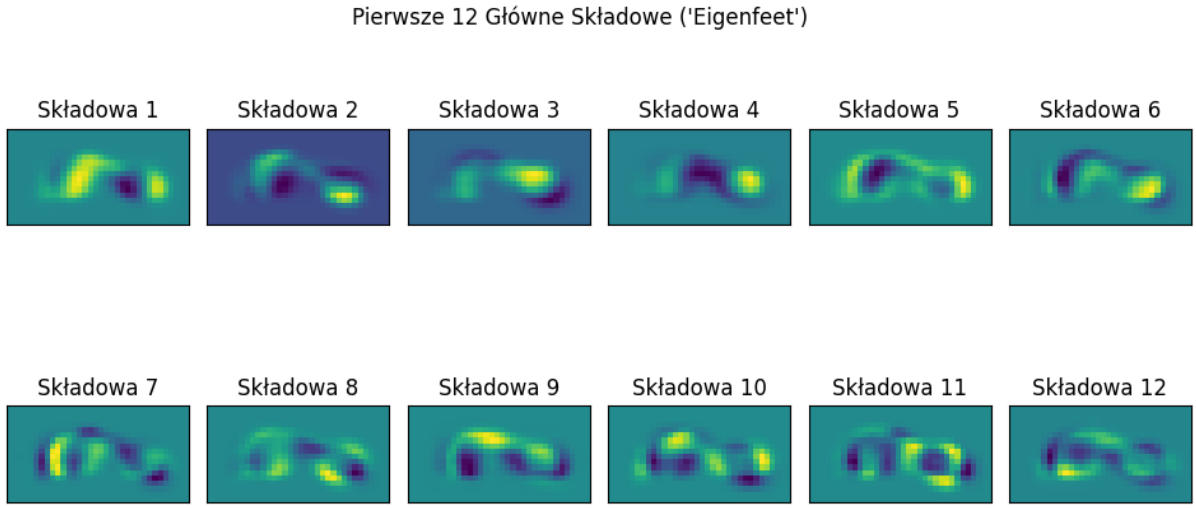
Rysunek 2: Skumulowana wyjaśniona wariancja w zależności od liczby głównych składowych.

Zgodnie z wykresem, aby wyjaśnić 95% całkowitej wariancji w zbiorze danych, wystarczy zachować jedynie **33** główne składowe. Pozwoliło to na znaczną redukcję wymiarowości danych.

Finalny model PCA został zbudowany z użyciem $k = 33$ składowych. Po transformacji, dane miały kształt `(6588, 33)`. Dokładny procent zachowanej wariancji wyniósł **95.13%**.

4.2 Wizualizacja Głównych Składowych ("Eigenfeet")

Główne składowe uzyskane z PCA, zwane wektorami własnymi macierzy kowariancji, można zwizualizować, przekształcając je z powrotem do oryginalnych wymiarów obrazu (18x34). Reprezentują one główne kierunki zmienności w danych. W analogii do "Eigenfaces", można je nazwać "Eigenfeet". Macierz komponentów `pca.components_` miała kształt `(33, 612)`. Rysunek 3 przedstawia wizualizację pierwszych 12 "Eigenfeet".

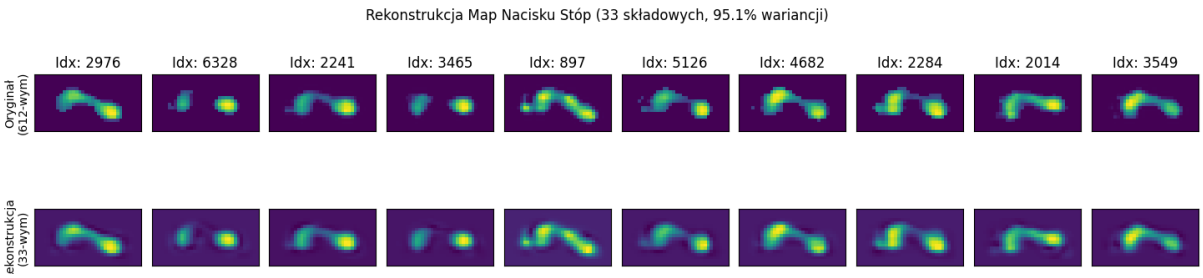


Rysunek 3: Wizualizacja pierwszych 12 głównych składowych ("Eigenfeet").

"Eigenfeet"pokazują wzorce nacisku, które najbardziej różnicują poszczególne próbki w zbiorze. Pierwsze składowe często kodują globalne cechy (np. ogólny kształt, podstawowy rozkład nacisku), podczas gdy kolejne składowe mogą reprezentować bardziej subtelne różnice.

4.3 Ocena Jakości Rekonstrukcji

Aby ocenić, ile informacji zostało utracone podczas redukcji wymiarowości do 33 składowych, zrekonstruowano obrazy stóp z danych o zredukowanej wymiarowości. Kształt danych zrekonstruowanych był zgodny z oryginałem: `(6588, 612)`. Rysunek 4 porównuje 10 losowo wybranych oryginalnych map nacisku z ich rekonstrukcjami.



Rysunek 4: Porównanie 10 losowych oryginalnych map nacisku stóp (górny rząd) z ich rekonstrukcjami na podstawie 33 głównych składowych (dolny rząd).

Wizualne porównanie pokazuje, że rekonstrukcje są bardzo zbliżone do oryginałów. Główne cechy rozkładu nacisku są dobrze zachowane, co potwierdza, że 33 główne składowe (wyjaśniające 95.13% wariancji) wystarczają do reprezentowania większości istotnych informacji zawartych w danych.

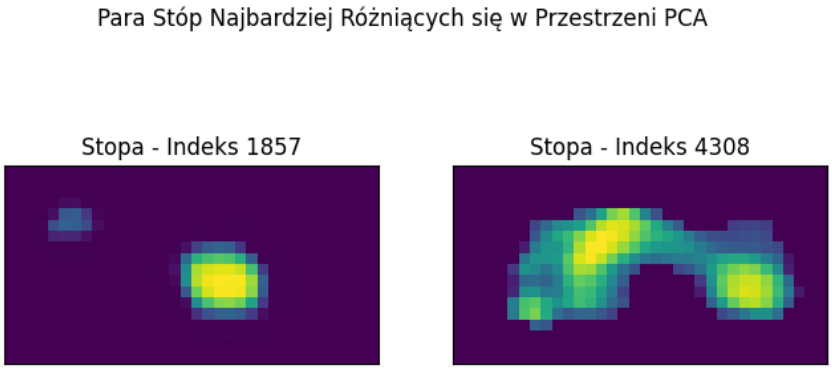
4.4 Identyfikacja Najbardziej Charakterystycznych Stóp

Analiza odległości euklidesowych między próbkami w przestrzeni PCA (33-wymiarowej) pozwoliła zidentyfikować parę map nacisku stóp, które najbardziej różnią się od siebie

pod względem cech uchwyconych przez główne składowe.

Porównanie wyników z funkcji `euclidean_distances` (sklearn) i `pdist` (scipy) wykazało, że obliczone macierze odległości **nie były identyczne** (wynik `np.allclose` to `False`). Może to wynikać z bardzo niewielkich różnic w precyzji obliczeń zmiennoprzecinkowych w implementacjach obu bibliotek. Jednakże, jak wskazuje dalsza analiza, różnice te nie wpłynęły na identyfikację pary o maksymalnej odległości.

Największa znaleziona odległość euklidesowa w przestrzeni PCA wyniosła **2153.83**. Odpowiadała ona parze stóp o oryginalnych indeksach: **1857** i **4308**. Wizualizacja tych dwóch map nacisku stóp przedstawiona jest na Rysunku 5.



Rysunek 5: Para map nacisku stóp (indeksy 1857 i 4308) o największej odległości euklidesowej w przestrzeni PCA (33 wymiary).

Te dwie mapy reprezentują ekstrema zmienności w zbiorze danych, uchwycone przez 33 najważniejsze główne składowe. Pokazują one skrajnie różne wzorce rozkładu nacisku.

5 Podsumowanie i Wnioski

5.1 Podsumowanie Analizy

W niniejszym sprawozdaniu przeprowadzono analizę zbioru danych map nacisku stóp przy użyciu metody PCA. Zbiór zawierał 6588 obrazów o rozdzielczości 18×34 pikseli. Dane zostały wstępnie przetworzone poprzez zmianę kształtu i spłaszczenie do formatu odpowiedniego dla PCA.

Zastosowanie PCA pozwoliło na:

1. **Zbadanie struktury wariancji:** Wykres skumulowanej wyjaśnionej wariancji pokazał, że znaczną część zmienności danych można wyjaśnić za pomocą relatywnie niewielkiej liczby głównych składowych. Wybrano 33 składowe, które wyjaśniają około 95.1% całkowitej wariancji.
2. **Redukcję wymiarowości:** Wymiar danych został zredukowany z 612 do 33, co znacząco upraszcza dane przy zachowaniu większości informacji.
3. **Wizualizację głównych wzorców ("Eigenfeet"):** Zwizualizowano pierwsze główne składowe, które reprezentują podstawowe wzorce zmienności nacisku stóp w analizowanym zbiorze.

4. **Ocenę rekonstrukcji:** Porównanie oryginalnych i zrekonstruowanych obrazów pokazało, że przy użyciu 33 składowych jakość rekonstrukcji jest wysoka, co potwierdza skuteczność redukcji wymiarowości.
5. **Identyfikację charakterystycznych stóp:** Znaleziono parę stóp (indeksy 1857 i 4308) o największej odległości euklidesowej w przestrzeni PCA, reprezentującą najbardziej różniące się od siebie przypadki w zbiorze pod względem uchwyconych przez PCA cech.

5.2 Wnioski i Wpływ Własności PCA

- **Skuteczność redukcji:** PCA okazało się skuteczną metodą redukcji wymiarowości dla zbioru map nacisku stóp, umożliwiając znaczące zmniejszenie liczby cech (z 612 do 33) przy minimalnej utracie informacji (zachowano >95% wariancji). Jest to kluczowa zaleta PCA w kontekście przetwarzania danych o wysokiej wymiarowości.
- **Interpretacja składowych:** "Eigenfeet"dostarczają wglądu w główne kierunki zmienności, ale ich bezpośrednia interpretacja fizjologiczna może być trudna (wada PCA - trudniejsza interpretowalność składowych). Mogą one reprezentować kombinacje różnych cech, takich jak ogólny kształt odcisku, rozkład nacisku między piętą a przodostopiem, czy specyficzne obszary wysokiego nacisku.
- **Identyfikacja ekstremów:** Analiza odległości w przestrzeni PCA pozwoliła na obiektywne znalezienie najbardziej nietypowych/różniących się stóp w zbiorze, co może być przydatne do dalszej analizy, identyfikacji outlierów lub zrozumienia zakresu zmienności w populacji. Obliczanie odległości w przestrzeni o zredukowanej wymiarowości pomaga uniknąć problemów związanych z "klątwą wymiarowości".
- **Ograniczenia liniowości:** PCA zakłada liniowe zależności między zmiennymi. Jeśli w danych istnieją złożone, nieliniowe struktury, PCA może nie być w stanie ich w pełni uchwycić. W przypadku map nacisku stóp, gdzie rozkład nacisku może być złożony, metody nieliniowej redukcji wymiarowości (np. t-SNE, UMAP) mogłyby potencjalnie dostarczyć dodatkowych informacji (np. do wizualizacji klastrów), choć PCA jest dobrym pierwszym krokiem i narzędziem do redukcji wymiarowości przed innymi analizami.
- **Zastosowanie w klasyfikacji:** Zredukowane dane (X_stopy_pca o wymiarze 33) mogą być użyte jako efektywne wejście do modeli klasyfikacyjnych (np. sieci neuronowych), co było wspomniane jako cel na kolejne zajęcia. Redukcja wymiarowości może przyspieszyć trenowanie modeli, zmniejszyć ryzyko przeuczenia i potencjalnie poprawić ich generalizację poprzez usunięcie szumu (składowych o małej wariancji).