

# Porównanie architektur Vision Transformer i CNN w zadaniu klasyfikacji zmian skórnych

Filip Rosiak - 151799, Eryk Stec - 152948, Kamil Krzywoń - 151776

PUT Poznań, 20 czerwca 2025

## Hipoteza badawcza

Vision Transformers osiągają lepszą dokładność niż tradycyjne CNN w klasyfikacji zdjęć dermatologicznych, szczególnie przy ograniczonych danych treningowych.

## Zbiór danych

**HAM10000:** 10,015 zdjęć dermatoskopowych, 7 klas zmian skórnych.

- ▶ **Melanoma (MEL)** - czerniak złośliwy
- ▶ **Nevus (NV)** - znamię barwnikowe
- ▶ **Keratosis (AK, BKL, DF)** - rogowacenie
- ▶ **Basal cell carcinoma (BCC)** - rak podstawnokomórkowy
- ▶ **Vascular lesions (VASC)** - zmiany naczyniowe

## METODY

### Modele:

- ▶ **ViT:** Vision Transformer Base Patch16-224 (86M parametrów)
- ▶ **ResNet50:** Residual Neural Network (23M parametrów)
- ▶ **EfficientNet-B0:** Efektywna architektura (4M parametrów)

**Eksperyment:** Testy z różnymi frakcjami danych (10%, 25%, 50%, 100%).

**Metryki:** Accuracy, F1-score, Confusion Matrix, mapy interpretability.

## WYNIKI EKSPERYMENTÓW

Tabela 1: Porównanie dokładności modeli

Model	10%	25%	50%	100%
ViT	<b>72.7%</b>	70.1%	74.7%	<b>86.2%</b>
ResNet50	70.3%	67.7%	<b>75.8%</b>	83.4%
EfficientNet	66.2%	<b>72.0%</b>	71.3%	84.3%

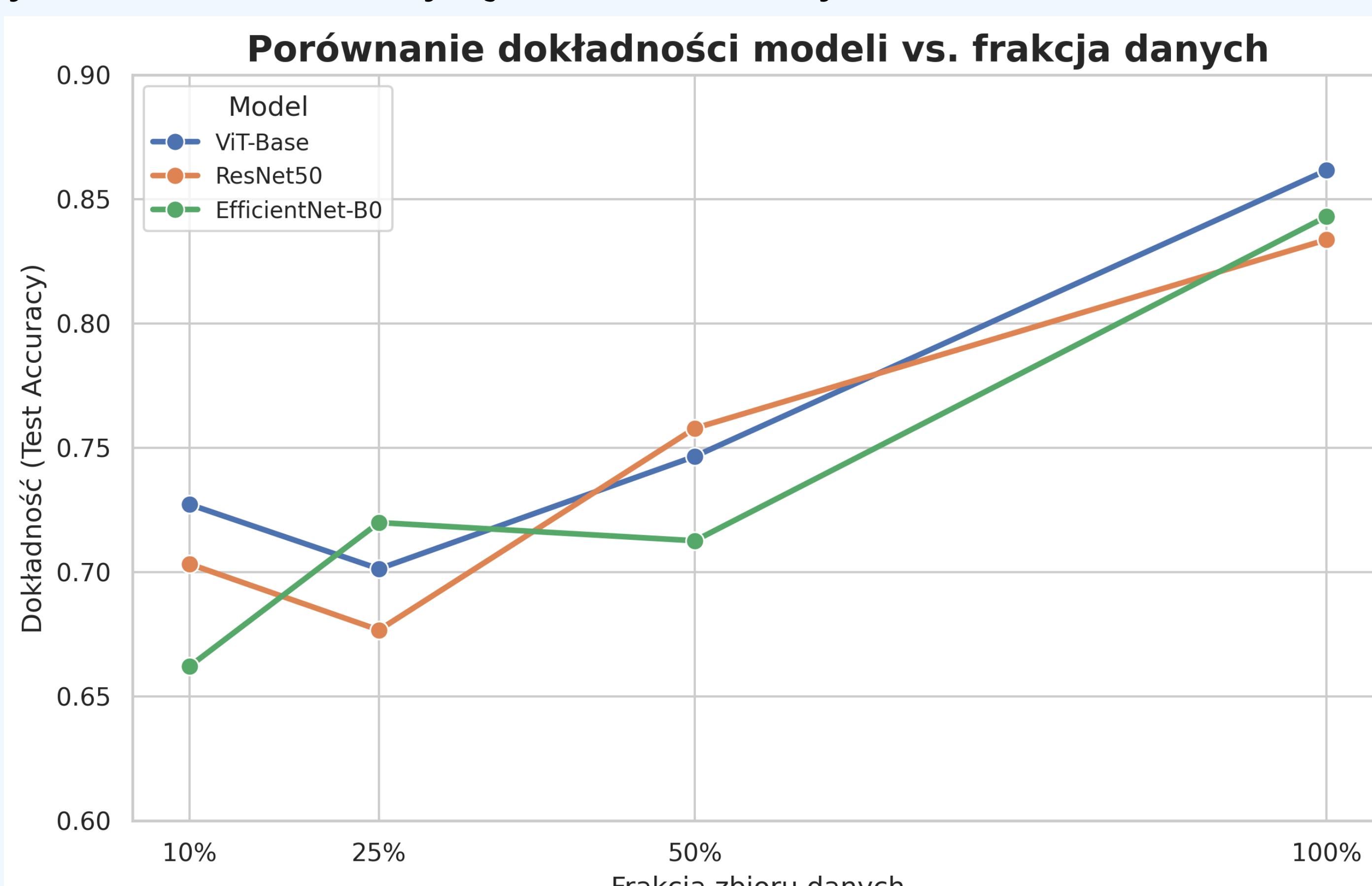
### Kluczowe obserwacje:

- ▶ ViT najlepszy przy małych (10%) i dużych (100%) zbiorach
- ▶ ResNet50 efektywny przy średnich ilościach danych (50%)
- ▶ EfficientNet najbardziej stabilny

### F1-Score (pełny zbiór):

- ▶ ViT: **86.5%**
- ▶ ResNet50: 83.3%
- ▶ EfficientNet: 84.4%

## Wykres 1: Porównanie wydajności vs ilość danych

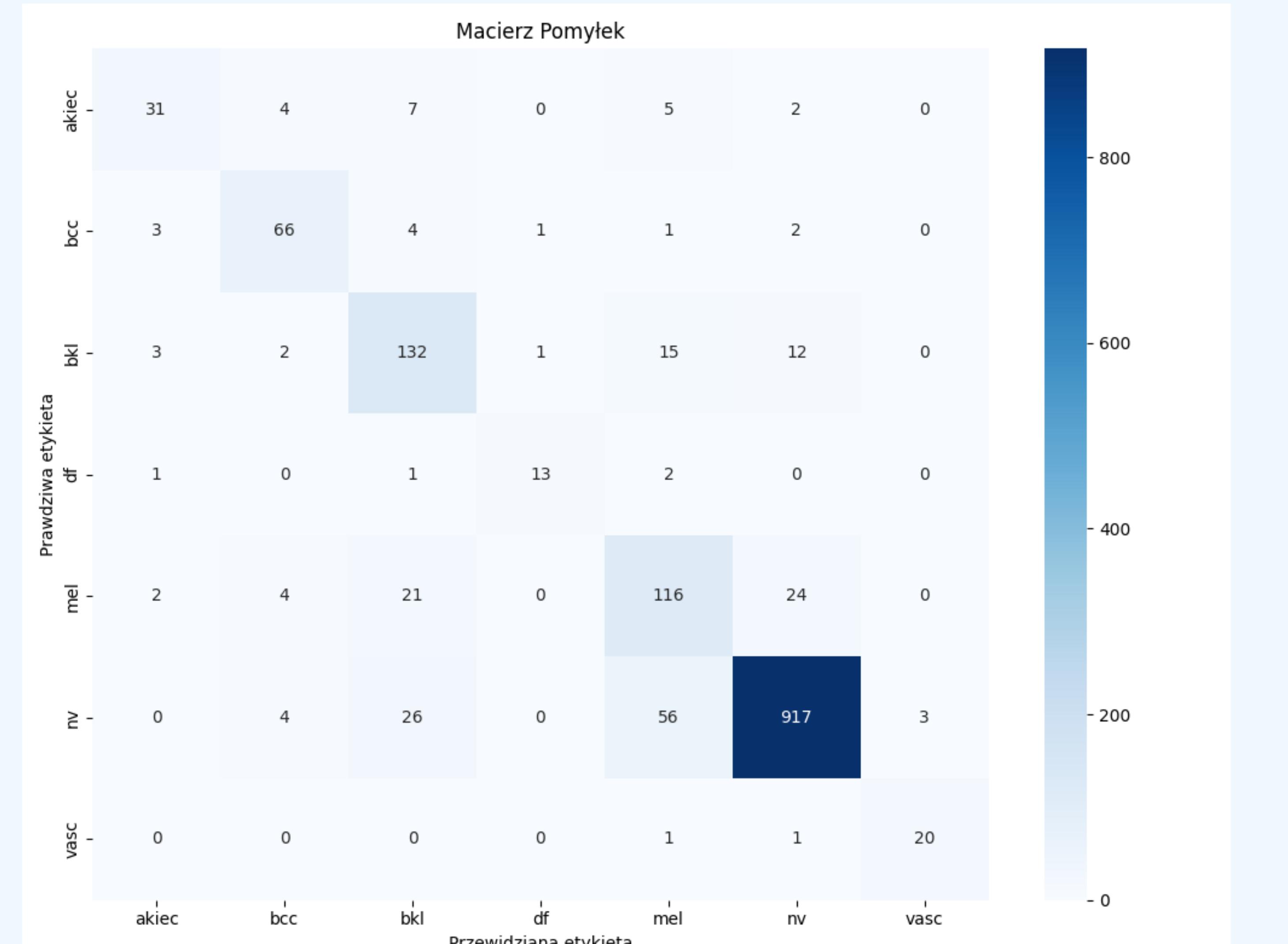


Wraz ze wzrostem liczby zdjęć do analizy zwiększa się dokładność modeli

## Bibliografia

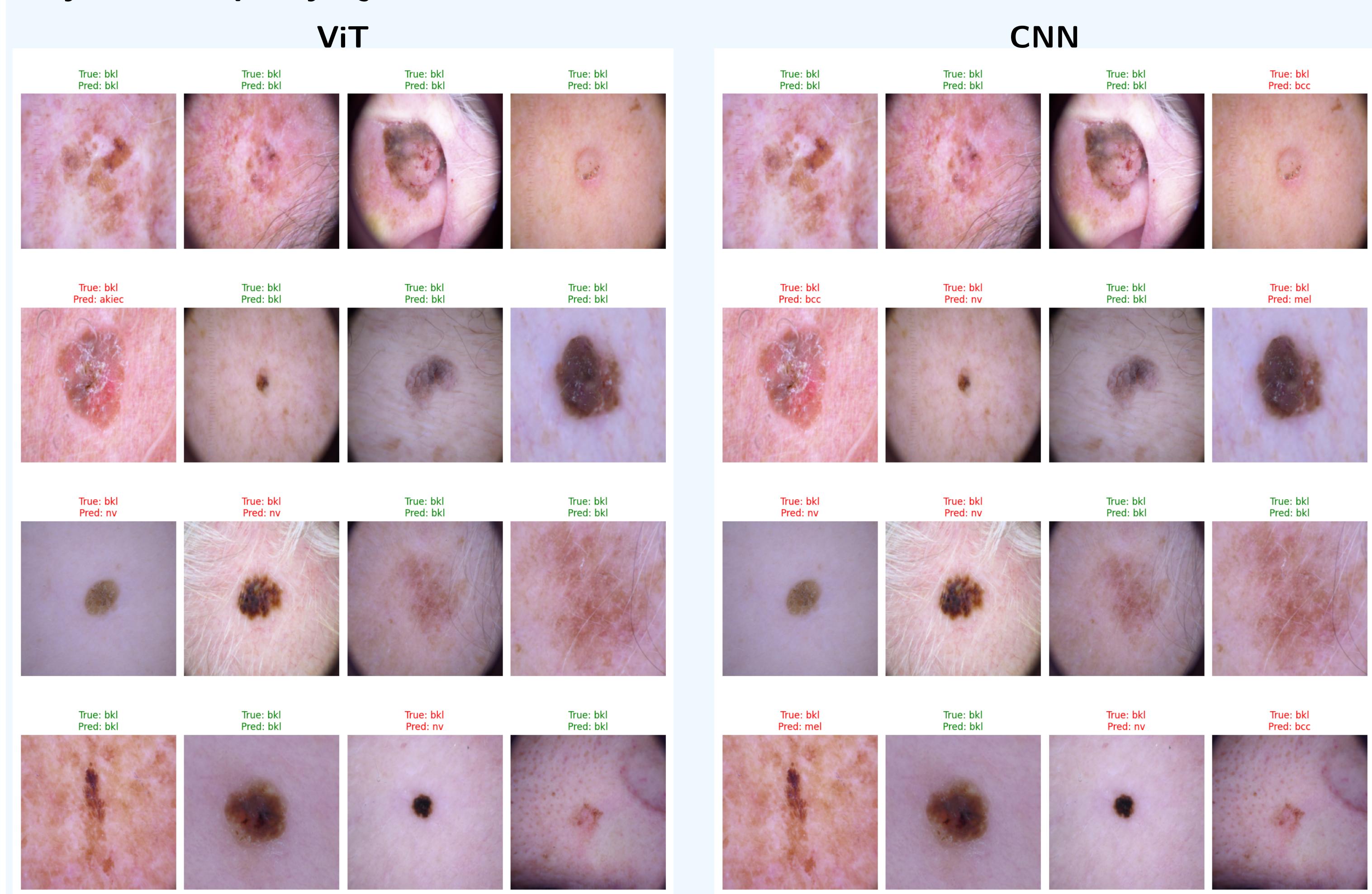
- ▶ Zbiór danych HAM10000  
<https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000>
- ▶ Vision Transformer model  
<https://huggingface.co/google/vit-base-patch16-224>
- ▶ ResNet50 model:  
<https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html>
- ▶ EfficientNetB0 model:  
[https://docs.pytorch.org/vision/main/models/generated/torchvision.models.efficientnet\\_b0.html](https://docs.pytorch.org/vision/main/models/generated/torchvision.models.efficientnet_b0.html)

## WIZUALIZACJE

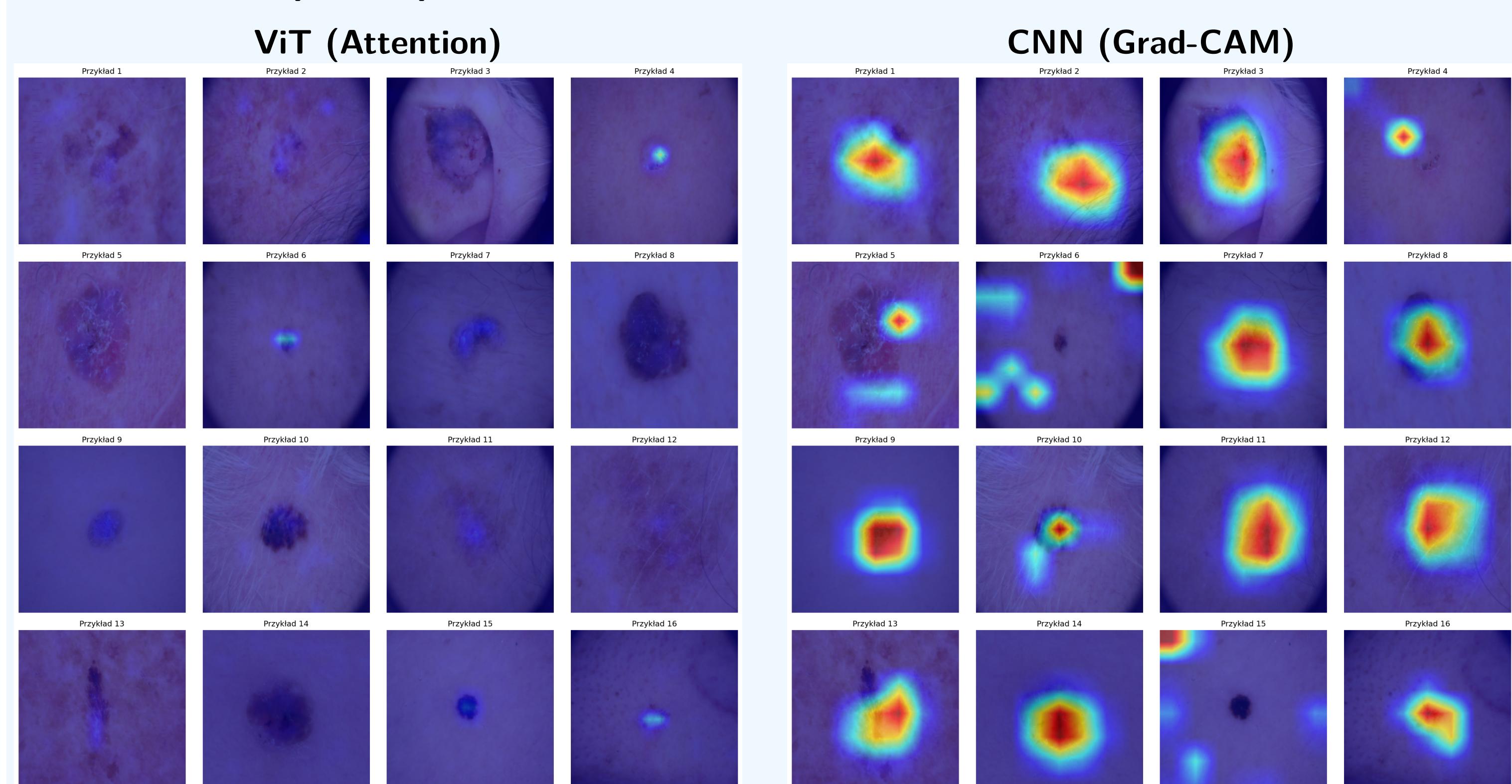


Macierz pomyłek dla ViT na pełnym zbiorze danych (100%)

### Przykładowe predykcje modeli



### Porównanie map interpretability



## Wnioski

Hipoteza, że Vision Transformers przewyższają CNN w klasyfikacji zmian skórnych, została częściowo potwierdzona. ViT osiągnął najwyższą dokładność przy 10% (72,7%) i 100% (86,2%) danych, ale ResNet50 był lepszy przy 50% (75,8%). EfficientNet wykazał największą stabilność. Wyniki sugerują, że ViT jest skuteczniejszy przy bardzo małych lub dużych zbiorach danych, podczas gdy CNN pozostają konkurencyjne w pośrednich scenariuszach.