# Dataset

**Skin Cancer MNIST - HAM10000**: This dataset consists of 10,000 dermatoscopic images of pigmented skin lesions across seven different diagnostic categories. The dataset was curated to train neural networks for automated diagnosis of pigmented skin lesions.

**Categories in dataset**:

- Actinic keratoses and intraepithelial carcinoma
- Basal cell carcinoma
- Benign keratosis-like lesions
- Dermatofibroma
- Melanoma
- Melanocytic nevi
- Vascular lesions

**Data attributes**:

- Image files in HAM10000_images
- Metadata including age, sex, localization, and diagnostic confirmation in HAM10000_metadata

# Data Preprocessing

- **Image Resizing**: Standardize all images to the same dimensions to ensure consistent feature extraction.
- **Color Normalization**: Standardize color representation to reduce variations in lighting and camera settings, which are common in dermatoscopic imagery.
- **Duplicates**: The dataset contains multiple images of the same lesion (with same lesion_id but different image_id)

# Classification Proposal

The project aims to build a model that can accurately classify dermatoscopic images into their corresponding skin lesion categories. This classification has significant real-world applications in assisting dermatologists with preliminary diagnosis and screening.

# Intended Features

- **Histogram of Oriented Gradients (HOG)**: To capture edge and texture patterns that are crucial for distinguishing between different types of skin lesions.
- **Color Histograms**: To quantify color distributions, as color is a key diagnostic feature in dermatology (e.g., melanomas often have specific color patterns).
- **Shape Descriptors**: To quantify border irregularity and asymmetry, which are important clinical indicators in skin cancer diagnosis.

# Classifiers

- **Convolutional Neural Networks (CNNs)**: This will be most likely the most accurate classifier, given that they are specifically designed for image processing tasks.
- **Random forests**: Provides robust classification with feature importance metrics that can help identify which image characteristics are most diagnostically relevant.

# Evaluation Metrics

Given the medical nature of this classification task and likely class imbalance, we'll focus on:

- Precision, recall, and F1-score per class
- Confusion matrix analysis
- ROC curves and AUC metrics
- Sensitivity and specificity, particularly for malignant vs. benign classifications