**Name : PRASHITHA S**

**USN : JUPG25MTECH22197**

# A DATA-DRIVEN ANALYSIS

**Domain:** Image Processing & Computer Vision

**Case Study:** Medical Image Analysis **(HAM10000 Dataset)**

## 1. Dataset Overview

- **Name**: HAM10000 ("Human Against Machine with 10000 training images")
- **Type**: Skin lesion dataset for dermatology research & image classification.
- **Size**:
  - ~10,015 dermatoscopic images (.jpg files in two folders)
  - Metadata file: HAM10000_metadata.csv
  - Additional preprocessed CSVs: hmnist_8_8_L.csv, hmnist_8_8_RGB.csv, hmnist_28_28_L.csv, hmnist_28_28_RGB.csv (downsampled image versions for ML models).

**Goal:** Build models that classify skin lesions into 7 diagnostic categories.

## 2. Metadata Attributes

From HAM10000_metadata.csv:

- lesion_id → Unique lesion identifier
- image_id → Image file name (maps to .jpg image)
- dx → Diagnosis label (target variable)
- dx_type → How the diagnosis was made (histo, consensus, follow_up, confocal)
- age → Patient age (some missing values)
- sex → Patient gender (male, female, unknown)
- localization → Body site of lesion (back, chest, lower extremity, etc.)

## 3. Class Distribution (Target Labels - dx)

The dataset has 7 categories of skin lesions:

- akiec → Actinic keratoses and intraepithelial carcinoma / Bowen's disease
- bcc → Basal cell carcinoma
- bkl → Benign keratosis-like lesions
- df → Dermatofibroma
- nv → Melanocytic nevi (most common)
- vasc → Vascular lesions
- mel → Melanoma

We can generate counts & percentages of each class to see imbalance (important for ML).

## 4. Demographic Information

- **Age**: Distribution of patients' ages.
- **Sex**: Ratio of male/female patients.
- **Localization**: Most common body sites affected.

## 5. Image Files

- Images are high-quality dermatoscopic images.
- Stored in two folders (HAM10000_images_part_1 and HAM10000_images_part_2).
- File names match with image_id in CSV.
- Pixel data is also available in downsampled CSV versions (8x8, 28x28, grayscale/RGB).

## 6. Potential Issues / Considerations

- **Imbalanced classes** (e.g., nv has many more samples than df or vasc).
- **Missing data** in age & sex.
- **Variability** in acquisition (lighting, body parts, skin tones).