Computer Vision: Classification of Skin Lesions from the HAM10000 Dataset

Team: Henri Antikainen, Bailey Forster, Tristan Freiberg
GitHub: https://github.com/tmfreiberg/HAM10000-skin-lesion-classification

Project Overview: Melanoma can affect anyone and early detection is a crucial factor affecting survival rates. Machine learning models could assist trained healthcare professionals in screening for skin cancer. The Human Against Machine 10000 (**HAM10000**) dataset contains images of 7,470 distinct skin lesions, each belonging to one of seven mutually-exclusive classes of skin lesion, including melanoma, as well as other cancerous types and benign types such as nevi (moles).

Our goal: Train a convolutional neural network (CNN) to accurately classify images of skin lesions using the HAM10000 dataset. The project includes a Streamlit app to test users' classification ability against our fine-tuned models.

Stakeholders:

- **Healthcare Workers:** confirm diagnosis or prompt further consideration.
- **Medical Students:** this project provides valuable hands-on experience in applying machine learning to healthcare challenges.
- **Public Awareness Campaigns:** initiatives like SunSmart" can benefit from improved skin lesion classification, aiding in public awareness and education.

Key Performance Indicators:

- Balanced Accuracy: BACC = (sensitivity + specificity)/2
- Sensitivity (True positive rate) = TP/(TP+FN)
- Specificity (True negative rate) = TN/(TN+FP)

Modeling Approach: We use the HAM10000 dataset to fine-tune two popular CNN architectures (ResNet and EfficientNet) to classify 5 categories of skin lesions

- **Cancerous (3 classes)**: melanoma (mel), basal cell carcinoma (bcc), and actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec)
- **Benign (2 classes):** nevi (nv), and other (contains the three remaining benign classes from the HAM10000 dataset)

Train-Test split: We use a stratified split based on the lesion ID to ensure that the proportions of each class are consistent in the training and test sets. In the case of multiple images per lesion, we either randomly select one image to keep, or assign all images of that lesion to the same set.
Balancing: The dataset is highly imbalanced, with cancerous lesions such as melanoma being extremely underrepresented (<10%) compared to benign nevi (>70%). We address this issue by undersampling larger classes and oversampling smaller classes, so that each class has 2000 images.
Augmentation: Random transformations such as random cropping, color jitter, and rotations are applied to augment the dataset and improve model generalization.

Results:
Model Performance: Balancing the dataset significantly improved the model performance. Our best results with Resnet18, using balanced data and random crop transformations, achieved 70% balanced accuracy. With similar model parameters, EfficientNet was able to achieve 73% balanced accuracy. The streamlit app allows users to test their ability to classify skin lesions as moles or melanoma, while competing against one of our fine-tuned machine learning models.

Future Work:
- Incorporate dropout regularization to mitigate overfitting.
- Address artifacts in images, such as markings from dermatologists, for improved classification accuracy.
- Include metadata (e.g. age, sex, localization of lesion) as model features.
- Test/finetune model on datasets with more diversity in ethnicity and skin tones.