Background

Dermoscopic lesion imaging, like many diagnostic medical exams, produces a long-tailed distribution of clinical findings; while a small subset of diseases are routinely observed, the vast majority of diseases are relatively rare. This poses a challenge for standard AIML methods, which exhibit bias toward the most common classes at the expense of the important, but rare, “tail” classes. Many existing methods have been proposed to tackle this specific type of imbalance, though only recently with attention to long-tailed medical image recognition problems.

Diagnosis on dermoscopic lesion images is a multi-class problem, as each image is associated with a single disease class. This dataset has 10, 015 skin images with 7 lesion classes. This dataset is characterized by an imbalance ratio (IR) of 58, where IR is defined as the ratio of the sample size of the largest majority class and that of the smallest minority class. Thus the larger the value of IR, the larger the imbalance extent.

Since most large-scale image classification benchmarks contain single-label images with a mostly balanced distribution of labels, many standard AIML methods fail to accommodate the class imbalance problem posed by the long-tailed nature of tasks like disease diagnosis on dermoscopic lesion images.

To develop a benchmark for long-tailed, multi-class medical image classification, we will use the HAM10000 Dataset that was acquired with a variety of dermatoscope types that are categorized into one of seven possible disease categories.

Goal

Submit automated predictions of disease classification within dermoscopic images.

Possible disease categories are:

1. Melanoma (MEL)
2. Melanocytic nevus (NV)
3. Basal cell carcinoma (BCC)
4. Actinic keratosis / Bowen’s disease (intraepithelial carcinoma) (AKIEC)
5. Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis) (BKL)
6. Dermatofibroma (DF)
7. Vascular lesion (VASC)

Input Data

The input data are dermoscopic lesion images in JPEG format.

All lesion images are named using the scheme ISIC_ .jpg , where is a 7-digit unique identifier. EXIF tags in the images have been removed; any remaining EXIF tags should not be relied upon to provide accurate metadata.

The lesion images come from the HAM10000 Dataset, and were acquired with a variety of dermatoscope types, from all anatomic sites (excluding mucosa and nails), from a historical sample of patients presented for skin cancer screening, from several different institutions. Images were collected with approval of the Ethics Review Committee of University of Queensland (Protocol-No. 2017001223) and Medical University of Vienna (Protocol-No. 1804/2017).

The distribution of disease states represent a modified "real world" setting whereby there are more benign lesions than malignant lesions, but an over-representation of malignancies.

Ground Truth Provenance

As detailed in the HAM10000 Dataset description, diagnosis ground truth were established by one of the following methods:

- Histopathology
- Reflectance confocal microscopy
- Lesion did not change during digital dermatoscopic follow up over two years with at least three images
- Consensus of at least three expert dermatologists from a single image

In all cases of malignancy, disease diagnoses were histopathologically confirmed.

KCDH Challenge Datasets:

Data	Description	Link
Training Data	10015 images	https://iitbacin-my.sharepoint.com/:f:/g/personal/30004952_iitb_ac_in/Eg-j-mjz0lxCpjxXrOFXyV8BLMdcgoa9kxGtuGCzisH8zg?e=ll1m1s

Additional Information	10015 entries grouping each lesion by image and diagnosis confirm type. A further explanation on this supplemental data can be found here	https://iitbacin-my.sharepoint.com/:x:/g/personal/30004952_iitb_ac_in/EfsqObcWWQhPq1bgOwjrrg0BwsjTXCEMhEflJQKZ-KxZXg?e=Qhhu2P
Training Ground Truth	One ground truth response CSV file (containing 1 header row and 10015 corresponding response rows).	https://iitbacin-my.sharepoint.com/:x:/g/personal/30004952_iitb_ac_in/EVY0CQA_Ls5AgVdkQrt5U_MBajSLyzjJbs5pCmDGQkvXdA?e=2MuhHe
Test Data	1512 images	https://iitbacin-my.sharepoint.com/:f:/g/personal/30004952_iitb_ac_in/EmBfcjPu3u5Pn9DU50_gHzMBJufaM8fLXG_rB0IVnQVR6Q?e=VBYkHv
Test Ground Truth	One ground truth response CSV file (containing 1 header row and 1512 corresponding response rows).	https://iitbacin-my.sharepoint.com/:x:/g/personal/30004952_iitb_ac_in/EX0IaQf6UvpHjZc3GN6vSoABWFeMgXWLvEsGm1YUhXPu3A?e=jZ4r3J

Additional Information:

Additional supplemental information is provided in the document: KCDH2024_Training_LesionGroupings.csv file.

This might be helpful when splitting the Training data for internal training / evaluation processes. Use of this data is optional, and no such data will be provided for the formal Validation and Test phases, where predictions have to be made based on single image data only.

The structure of this supplemental information is as follows:

- For each image in the Part 3 Training set (labeled in this CSV as the column "image"), there is a lesion identifier (the "lesion_id" column) and a diagnosis confirm type methodology (the "diagnosis_confirm_type" column).

- Images with the same lesion identifier value show the same primary lesion on a patient, though the images may be taken at different camera positions, lighting conditions, and points in time.

- Images with a more rigorous diagnosis confirm type methodology are typically more difficult cases for human expert clinicians to evaluate, particularly when the images' ultimate diagnosis is benign. In ascending order of rigorousness as applied to cases of the present dataset, the diagnosis confirm type methodologies are "single image expert consensus", "serial imaging showing no change", "confocal microscopy with consensus dermoscopy", "histopathology".

- So, for example, an image with a diagnosis of "NV" ("Melanocytic nevus") which was confirmed by "histopathology" would typically be a more ambiguous or difficult case for human experts than a similar "NV" ("Melanocytic nevus") image confirmed by "single image expert consensus" only, as the former image required a more invasive procedure to diagnose.

If this data is chosen to be utilized in the algorithm training process, it is the responsibility of the user to figure out how to best incorporate it.

Evaluation

Goal Metric

Predicted responses are scored using a normalized multi-class accuracy metric (balanced across categories). Tied positions will be broken using the area under the receiver operating characteristic curve (AUC) metric.

Other Metrics

Participants will be ranked based only on the multiclass accuracy metric. However, for scientific completeness, predicted responses will also have the following metrics computed (comparing prediction vs. ground truth) for each image:

Individual Category Metrics

- sensitivity
- specificity
- accuracy
- area under the receiver operating characteristic curve (AUC)
- mean average precision

- F1 score
- average AUC across all diagnoses
- positive predictive value (PPV)
- negative predictive value (NPV)

Supporting papers

(1) Robust Asymmetric Loss for Multi-Label Long-Tailed Learning (Paper / Code)

(2) Focal Loss for Dense Object Detection (Paper / Code)

(3) Asymmetric Loss For Multi-Label Classification (Paper / Code)

(4) Simple and Robust Loss Design for Multi-Label Learning with Missing Labels (Paper / Code)