**SEIS 764-01: Artificial Intelligence**

**Title: Skin Cancer Classification**

Project Proposal

Due Date: 1st May

Submitted by:

Mohammed Hesham Ahmed (ahme3785@stthomas.edu)

Sravani Beegala (beeg8557@stthomas.edu)

Oge Edward (edwa9954@stthomas.edu)

Christopher Markey (chris.a.markey@seagate.com)

Shubha Shubha (shub6690@stthomas.edu)

**1. Description of data source and web link(s).**

The HAM10000 dataset is a large collection of multi-sources dermatoscopic images of common pigmented skin lesions. Dermatoscopic images are collected from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all-important diagnostic categories in the realm of pigmented lesions.

**Dataset Web Link:** <https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000>

**2. Size/# of records of the dataset or files.**

The dataset includes lesions with 10015 images, which can be tracked by the lesion\_id-column within the **HAM10000\_metadata** file. The metadata file also has details of each image associated with image\_id such as lesion\_id, dx, dx\_type, age, sex, localization.

**3. # of attributes of the dataset.**

Dataset has 7 attributes along with image associated with each image id.

1. lesion\_id
2. image\_id
3. dx
4. dx\_type
5. age
6. sex
7. localization

**4. Description of each attribute.**

The attributes are basically the images for each class:

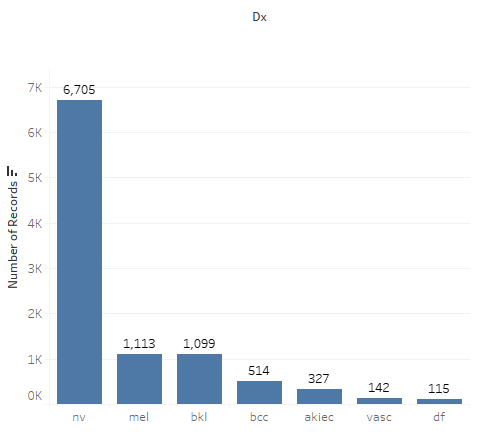
1. **AKIEC**: Actinic Keratoses (Solar Keratoses) and intraepithelial Carcinoma (Bowen’s disease) are common non-invasive, variants of squamous cell carcinoma that can be treated locally without surgery. Actinic keratoses are more common on the face and Bowen’s disease is more common on other body sites. Because both types are induced by UV-light the surrounding skin is usually typified by severe sun damaged except in cases of Bowen’s disease that are caused by human papilloma virus infection and not by UV
2. **BCC**: Basal cell carcinoma is a common variant of epithelial skin cancer that rarely metastasizes but grows destructively if untreated. It appears in different morphologic variants (flat, nodular, pigmented, cystic, etc.)
3. **BKL**: "Benign keratosis" is a generic class that includes seborrheic keratoses ("senile wart"), solar lentigo - which can be regarded a flat variant of seborrheic keratosis - and lichen-planus like keratoses (LPLK), which corresponds to a seborrheic keratosis or a solar lentigo with inflammation and regression.
4. **DF**: Dermatofibroma is a benign skin lesion regarded as either a benign proliferation or an inflammatory reaction to minimal trauma. It is brown often showing a central zone of fibrosis dermatoscopically.
5. **MEL**: Melanoma is a malignant neoplasm derived from melanocytes that may appear in different variants. If excised in an early stage, it can be cured by simple surgical excision. Melanomas can be invasive or non-invasive (in situ). We included all variants of melanoma including melanoma in situ, but did exclude non-pigmented, subungual, ocular or mucosal melanoma
6. **NY:** Melanocytic nevi are benign neoplasms of melanocytes and appear in a myriad of variants
7. **VASC:** Vascular skin lesions in the dataset range from cherry angiomas to angiokeratomas and pyogenic granulomas. Hemorrhage is also included in this category.

The metadata file has below attributes:

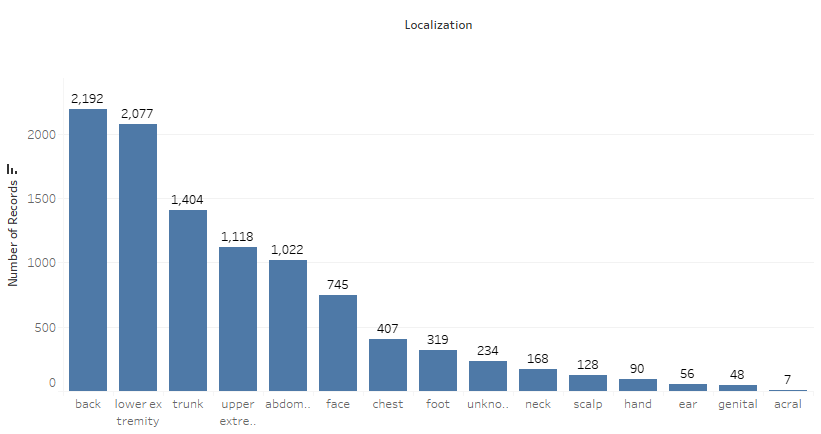
|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| lesion\_id | Lesion Id Number |
| image\_id | Image Id Number |
| dx | Lesion Class label |
| Dx\_Type | Lesion Confirmation method type |
| Age | Age number |
| Sex | Gender |
| Localization | Location of Lesion |

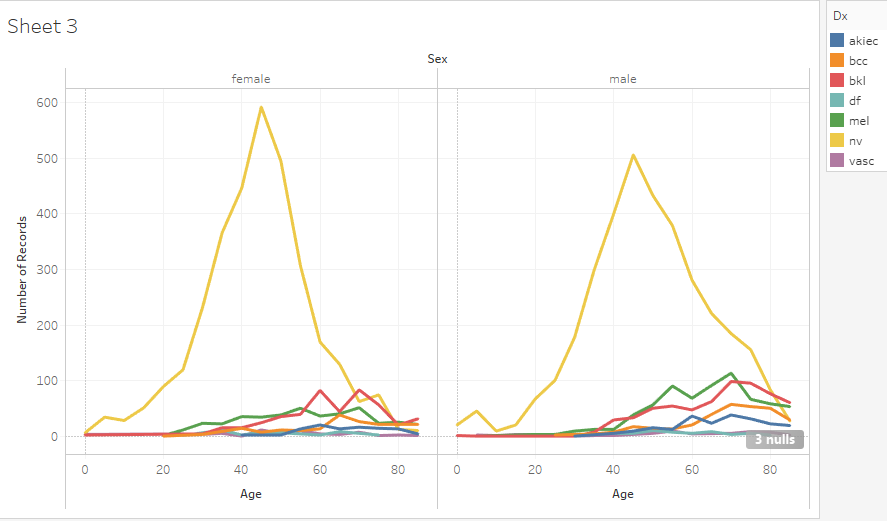
**5. Some general statistics of the dataset.**

The dataset has following distributions for the 7 classes:

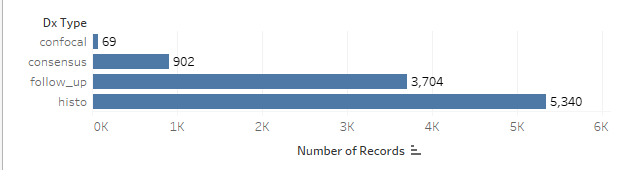


From the distribution of the data we can see the data set looks unbalances for few classes. In order to handle this problem, we propose to do the image augmentation for the classes with a smaller number of images to make our dataset more balanced

Histogram showing the localization of the skin cancer in human body. It seems the dataset has more images for lesion at back, trunk, abdomen and face then neck, scalp, hand and other body parts.



Line chart to compare the distribution of different skin cancer type for male and female with age.



More than 50% of lesions are confirmed through histopathology (histo), the ground truth for the rest of the cases is either follow-up examination (follow\_up), expert consensus (consensus), or confirmation by in-vivo confocal microscopy (confocal).

**6. Tools / methods you plan to use in your study.**

We are planning to use Google colab, anaconda, Kaggle, available image processing and deep learning libraries from Keras, Tensorflow, CNN, DNN, Image augmentation and GPU for computations.

**7. Exactly what problems/questions your team plans to predict / study.**

We want to do Skin Lesion Analysis and build a CNN model based on Lesion Segmentation and Disease Classification for different types of skin cancer with emphasis towards Melanoma Detection (giving more weight to it). We want to compare the F1 score and accuracy of different transfer learning models such as ResNet34/50, MobileNet, VGG16/19 and InceptionResnetV2/V3.