CSC490 Report – Group name: Team 8 – Skin Cancer Challengers

Team members: Arthur Alexandro Soenarto, Gabriel El Haddad, Xiaoning Wang, Syed Taha Ali

Name of dataset: Skin Cancer MNIST: HAM10000

Paper reference(s):

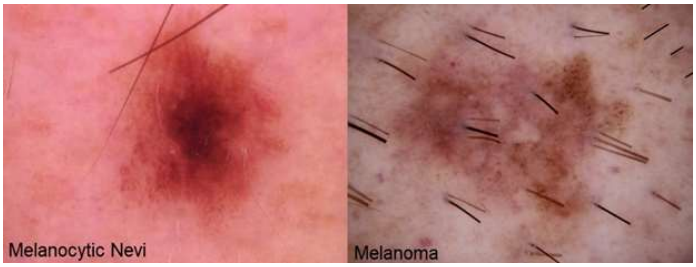
[Disease Classification based on Dermoscopic Skin Images Using Convolutional Neural Network in Teledermatology System](https://ieeexplore.ieee.org/abstract/document/8973303)

[Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)](https://arxiv.org/abs/1902.03368)

Location of the dataset: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

Background:

Skin Cancer is one of the most common cancers in North America. The most common cause is from overexposure to ultraviolet rays from the sun. It involves the growth of abnormal cells in the outermost skin layer (called the epidermis), which can form malignant tumors if not treated early. Since it grows in the outermost layer, this property makes skin cancer easily detectable and extremely relevant to camera-based machine learning applications, which is the motivation for our project.



Skin Cancer Mnist: HAM10000. (n.d.). kaggle. Retrieved October 3, 2022, from https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000.

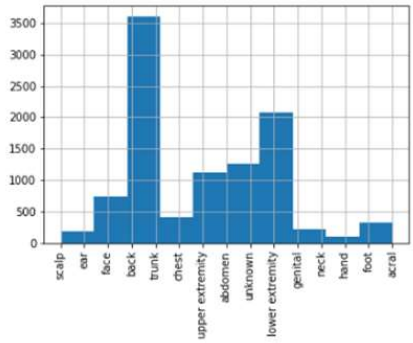
Reasons, why we chose this specific dataset:

The Skin Cancer MNIST: HAM10000 dataset contains dermatoscopic (examination of skin lesions) images from different populations. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes.

The dataset labels are very reliable as more than 50% of lesions are confirmed through histopathology. We are excited to use this dataset so we can combine the different skills we've discussed in class as far as segmentation, attribute detection, and disease classification.

The most important reason we chose this dataset is because it contains variable data that is representative of the real world. For example, the dataset contains image samples from over 15 different locations across the patient's bodies, image samples include both genders, and contain different skin shades. Having a dataset with various types of samples is important because it will allow the model to generalize to the real world and prevent it from being biased towards a specific sample of the population.

```
x = data['localization'].hist(xrot=90)
```




```
print(data['localization'].value_counts())
```

```
back               2192
lower extremity    2077
trunk              1404
upper extremity    1118
abdomen            1022
face                745
chest               407
foot                319
unknown             234
neck                168
scalp               128
hand                 90
ear                  56
genital              48
acral                 7
Name: localization, dtype: int64
```

Who started to access the dataset: Syed Taha Ali, Arthur

Who set up the source code repo: Arthur

Distribution of responsibilities in the group:

Arthur: - Also work on segmentation for model, Analyze dataset	Gabriel El Haddad: - Work on segmentation for the model
Xiaoning: - Work on segmentation for model, Find the relevant models, Setting up the google colab	Syed Taha Ali: - Data augmentation & analysis, Work on lesion classification model

Distribution of tasks for this report:

Arthur: - Setting up source code repo, download data	Gabriel El Haddad: - Explain reasons for choosing dataset, writing out the report in pdf format
Xiaoning: - Looking for methods, models and novel ideas, find connected papers, read the original competition paper to distribute the tasks	Syed Taha Ali: - Created figures for skin lesion localization in the dataset on Jupyter Notebook, added reasons for choosing dataset, described the problem we aim to solve with this project