**CMPSC 496 Capstone Weekly Report**

**GitHub**

**GitHub Repository:** [**https://github.com/jinyoonok2/Skin-Cancer-Detection-Capstone**](https://github.com/jinyoonok2/Skin-Cancer-Detection-Capstone)

**Dataset**

1. **ISIC 2018:** [**https://challenge.isic-archive.com/data/#2018**](https://challenge.isic-archive.com/data/#2018)

* The HAM10000 dataset is included in ISIC 2018 and is also a part of the ISIC 2019 dataset.
* A portion of this dataset was manually labeled using Roboflow.
* This manually labeled data was employed to train a model capable of automatic and comprehensive self-labeling.
* The trained model was subsequently utilized to assist in instance segmentation within the combined dataset.

1. **Combined Dataset:** only training proportion of each was used (50000+)

* ISIC 2019: <https://challenge.isic-archive.com/data/#2019>
* ISIC 2020: <https://challenge.isic-archive.com/data/#2020>

**Meeting**

1. **Time and place of the meeting:** 2:00 PM Jan/12/2024
2. **Who was present:** Professor Md Faisal Kabir, Jinyoon Kim, Tianjie Chen
3. **What are their feedbacks/ suggestions:**

* Review the classification results to identify specific instances of misclassification, such as predictions where one class is mistakenly identified as another.
* Elaborate on the purpose and advantages of each model or method being employed.
* Detail the preprocessing steps applied to the dataset to enhance model performance.
* Develop a web-based application prototype, building upon the initial model prototype.

1. **What have you discussed with the faculty advisor:** check the first report below.

For the more progressed and detailed weekly report, you can check the link at the bottom.

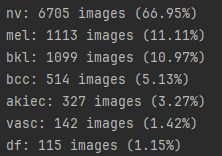
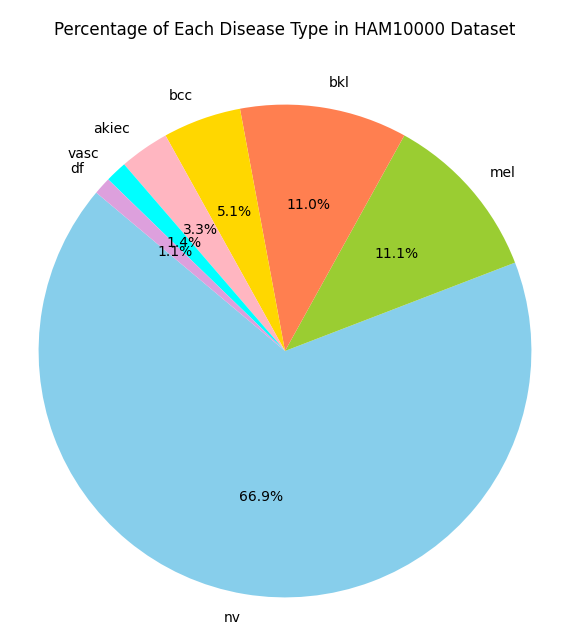
**First Report**

1. **The model is trained on ISIC 2018 dataset. The information of the model trained on ISIC 2018 (HAM10000) is like the following:**

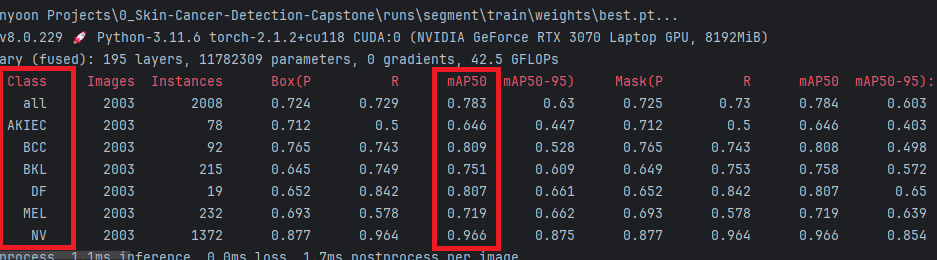
* Model Type: YOLOv8s-seg.pt (segmentation model)
* Class Names:

**AKIEC:** Actinic keratoses and intraepithelial carcinoma / Bowen's disease.  
**BCC:** Basal cell carcinoma.  
**BKL:** Benign keratosis-like lesions, including solar lentigines, seborrheic keratoses, and lichen-planus like keratoses.  
**DF:** Dermatofibroma.  
**MEL:** Melanoma.  
**NV:** Melanocytic nevi.  
**VASC:** Vascular lesions, including angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage.

* Dataset Distribution

* Performance of the YOLOv8 pretrained model trained on this dataset (20 epochs):



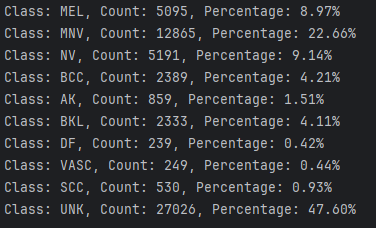
1. **Next, the segmentation model that is trained on ISIC 2018 dataset was used to assist most of the labeling process of the combined dataset (ISIC 2019 + ISIC 2020).**

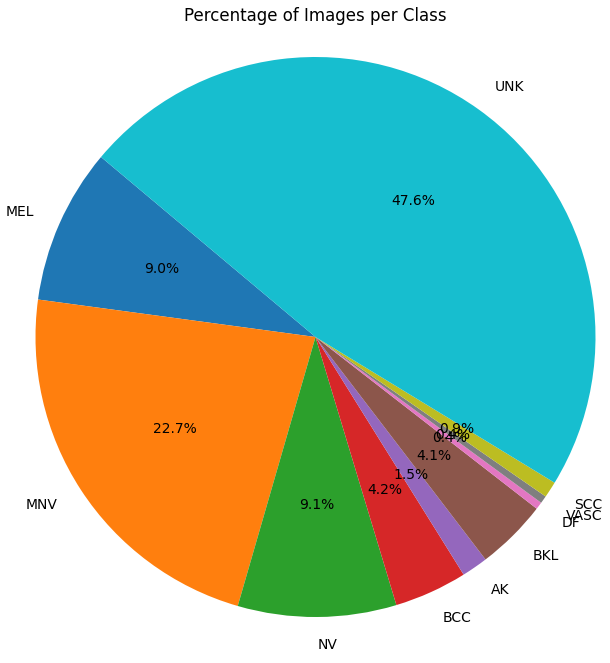
* Combined dataset is later manually processed through Roboflow computer vision platform to correct missing or crucially misinterpreted labels.
* Dataset is oriented and resized (640x640, 256x256) then downloaded to local to perform experiments
* Class correction, data splitting, data duplication for the preparation of the data augmentation, and minor processes were done locally.
* New class Names of the combined dataset:

**MEL:** Melanoma.

**MNV:** Melanocytic nevus.  
**NV:** Nevus.  
**BCC:** Basal cell carcinoma.  
**AK:** Actinic keratoses and intraepithelial carcinoma / Bowen's disease.  
**BKL:** Benign keratosis-like lesions, including solar lentigines, seborrheic keratoses, and lichen-planus like keratoses.  
**DF:** Dermatofibroma.  
**VASC:** Vascular lesions, including angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage.  
**SCC:** Squamous cell carcinoma  
**UNK:** All unknown cases from ISIC 2020 are diagnosed as **benign**.

* Classes with a very small population of data were combined into larger, more inclusive categories to reduce confusion.
* Dataset distributions for the combined dataset:





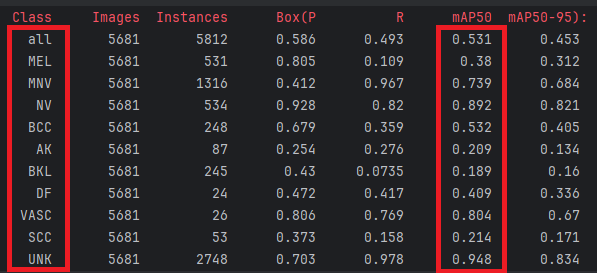
* For data duplication, a magnification setting is applied to each class. For instance, if the setting is 8, each data point in the class should be duplicated 7 additional times.

**magnification factors = {**  
 **'MEL': 2, 'MNV': 1, 'NV': 2, 'BCC': 2, 'AK': 4,**  
 **'BKL': 2, 'DF': 8, 'VASC': 8, 'SCC': 4, 'UNK': 1**  
**}**

* The duplication process is applied exclusively to the training split of the dataset. After this, the customized dataset class, with random augmentation, is passed through the data loader to initiate the training phase.

1. **First experiment with 20 epochs**

Model Type: YOLOv8s.pt (detection model)



1. **Investigate methods for integrating a transformer model into our system.**
2. **Consider utilizing an online service that offers powerful GPU instances to accommodate the significant computational demands of transformer models (vast.ai).**

**Further reports up to the current week and its progress can be found in the weekly report words file at the below link (only related professors and students can access):**

[**Weekly Report for CMPSC 488 Capstone Project Link**](https://pennstateoffice365-my.sharepoint.com/:w:/r/personal/juk481_psu_edu/_layouts/15/Doc.aspx?sourcedoc=%7B7160391E-7962-42A4-B68D-B7F3686CD747%7D&file=CMPSC%20488%20Capstone%20Project%20Weekly%20Report.docx&action=default&mobileredirect=true&DefaultItemOpen=1&login_hint=juk481%40psu.edu&ct=1705714744313&wdOrigin=OFFICECOM-WEB.START.EDGEWORTH&cid=e84eb1c5-f799-4a51-8158-fa54018b8e4b&wdPreviousSessionSrc=HarmonyWeb&wdPreviousSession=4109336f-307b-405f-b2fe-848c949f61d7)