Hi Yamini didi, He, and Prashant,

Here is the link to the CMIL-project website that shows our progress in the project, and we update it every week: [https://arvindkrishna87.github.io/STAT390\_project/Links to an external site.](https://arvindkrishna87.github.io/STAT390_project/)

Currently, while Yamini didi’s team labels the data, we are working on the following things:

1. **Patching algorithm** – Developing an efficient patching algorithm that will create patches from the epithelium-annotated and CMIL class-labelled data.
   1. We will first execute an algorithm on all the 107 cases that will create patches without any constraints on patch size, such that the patches collectively cover almost the entire epithelium, and each patch covers the width of the epithelium.
   2. Then, we will make a histogram of the patch sizes to see the modality of the patch-size distribution. Suppose most of the patch sizes are clustered around 3-4 values. Then, we’ll decide to train our classification model with those 3-4 distinct patch-sizes. *Note that we need to have a limited number of patch sizes to keep our ML model simple, and avoid complications.*
2. **Aspects of our ML algorithm** – We are surveying on different aspects of the CNN classification model that will classify our patches to benign / low-grade / high-grade. Some of the aspects are as follows:
   1. Should the patch size be in powers of 2 – 16 x 16, 32 x 32, 256 x 256, and so on, or can it be whatever fits best our data? How does it impact computational efficiency and model accuracy?
   2. If we make more patches to cover more of the epithelium, then we get a higher overlap – how to tradeoff between patch overlap & %area of epithelium covered. If we cover more of the epithelium, it also leads to more background area in patches – is that ok, or does if effect computational speed or model accuracy?

3. We are also surveying on some **medical aspects** of the problem:

* 1. Should we train our ML model to identify some patterns such as cysts, cellular formation (cells line up neatly or not), nuclear features (nuclei similar in size / shape)), maturation (gradual transition in cell size with depth), and follow it up with a decision tree to classify the lesion, or should we train the model to directly classify the lesion? Identifying specific patterns may be easier for the model, instead of directly classifying the lesion.
  2. There are some other ideas like comparing a patch with the rest of the patches of the same tissue slice, and if the patch is too different from other patches of the same tissue, then it indicates tendency towards melanoma *(UDI – Ugly Duck Index).*

Looking forward to our meeting next week!

Regards,

On Behalf of STAT390,

Krish.