User Story

Who is the user? What do they want to do with the tool? What needs and desires do they want for the tool? What is their skill level?

Hanson is a professional doctor. He wants to analyze images of breast tissue for his patients to diagnose breast cancer, specifically for early-stage prognosis. To be effective, the software should be easy to use for a non-data science user and produce quantitative results. The results should be reproducible between different doctors and/or sample preparation techniques.

I am a lab technician working with images of breast cancer tissue. I need a user-friendly platform where I can upload large datasets of images from my research and apply machine learning algorithms for classification.

As a data science researcher in the Mittal lab, I want to use machine learning models to analyze stained images of breast cancer tissue and distinguish between various immune cell types. I want the software to be well-documented and organized such that additional features can be implemented in future iterations.

Use Case

- Analyze tissue images/scans
- Diagnose breast cancer
- Advice for treatment options based on prognosis
    - Input: tissue images/scans
    - Output: some kind of quantitative score indicating how far along the tumor is

Components

***Name*: Pre-processing**
What it does: Takes the image file of the whole slide image (WSI) and separates it into smaller patches which are then sampled and used for creating the superpatch.
Inputs: WSIs, usually .svs or .ndpi
Outputs: Patches + superpatch (.tif) of the original image(s) that will be used to create the clustering model.
Side effects:

- Loss of data (when dividing WSI into patches)

Subcomponents:

- Differentiating background vs. tissue
- Converting image RGB values to grey scale values
- Sampling patches to put into the superpatch

***Name*: Cluster Processing**
What it does: generate human and computer readable data for analysis based on the clusters generated via the clustering model

Inputs: Original WSIs; clustered images
Outputs: Images which represent the original image and relevant overlays of the determined clusters. Generates multiple files: (*note: this module is imported into the seg.py file*)

- TILs overlayed on the original H&E patch
- binary segmentation masks of each cluster
- individual clusters overlayed on the original patch
- image of all the clusters
- CSV file containing contour information of each TIL segmented from the patch

Side effects:

- Incorrect identification of clusters

Subcomponents:

- Creating the image masks, contours, etc.
- Overlaying the colored masks, contours, etc. with the original image

Name: **Image Segmentation**
What it does: Applies a clustering model to patches to segment TILs in H&E stained images.
Inputs: Pre-processed image of breast cancer tumor
Outputs: Calls upon the image_postprocessing functions from tilseg.cluster_processing create readable images containing the clusters based on the optimized model.
Side effects:

- Inconclusive/inaccurate image result
- Not enough/too many clusters
- Disagreement between doctor & diagnosis

Subcomponents:

- Fitting the model to the superpatch + generating a score
- Applying the model to the other patches

<u>Things to Consider</u>

- How long will the software take to run each image?
- Accuracy?

<u>Possible Users</u>

- Researchers
- People updating/managing the software
- Lab technicians
- Doctors
    - Someone who has a lot vs. little experience with diagnosis