**EEL 5934 Project Report**

**Xin Ma, Yuyang Zhang**

**Introduction**

Our project goal is to create two pipelines. One of the pipelines uses lung's histology image and CODEX image to mark different cell types and map them to the histology image. And we also create another pipeline that accepts a kidney's histology image containing glomeruli as input and outputs a kidney image with different nuclei types marked with color. The two pipelines have similar structures. These two pipelines would help researchers to more easily classify cells in large images.

**Implementation**

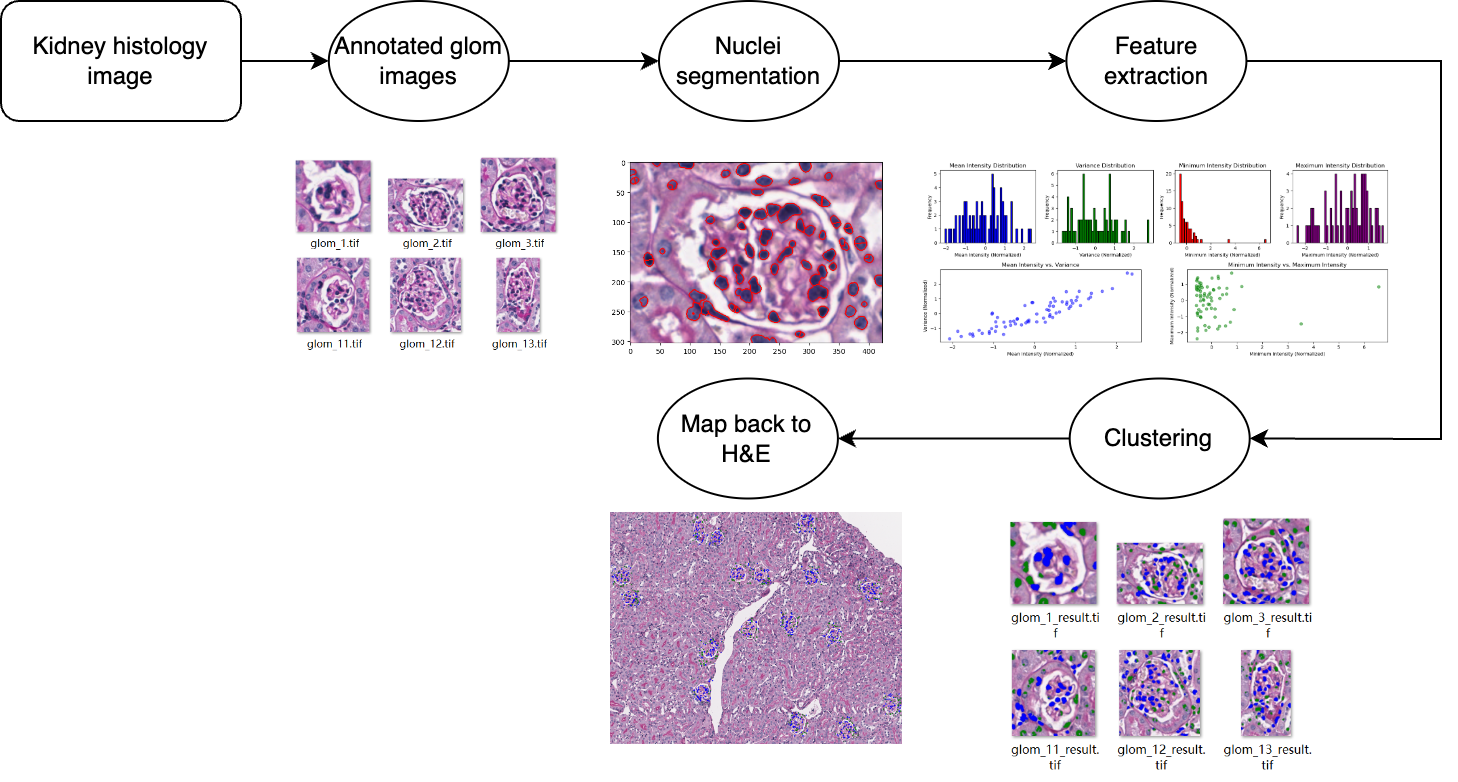
* Glomeruli pipeline (Fig. 1)

First, we use the ImageScope tool to annotate the area where glomeruli are located in the kidney image (.svs) and save the position to a .xml file. We can parse the .xml file and extract the image containing the glomerulus at its corresponding location. From this we obtain 104 glomerulus images.

For each color image, we apply color deconvolution to convert it to the HED color space and only retain the first channel (Hematoxylin). Then we use the Otsu method to convert the H channel to a binary image. We apply morphological opening operations to remove small objects and smooth edges using disc-shaped structuring elements. Next, we need to compute the distance transform on the binary mask and detect local maxima to apply watershed segmentation to separate the touching nuclei.

Through the above method, we obtain the segmentation results of each glomerulus image. Following segmentation of nuclei, we proceed with feature extraction on the segmented image. Subsequently, K-means clustering is applied to generate cluster labels, which are then assigned to each core region. For feature extraction, we calculate the mean intensity, variance, minimum intensity, and maximum intensity for each segmented nuclei and apply Z-score normalization. The features selected are statistical features rather than morphological features because we consider that the shape of the nuclei would be affected by the nuclei segmentation and cannot express its original morphology. We let K-means clustering generate two clusters, performed on a feature matrix consisting of normalized features. Cluster color is determined based on the intensity, with blue representing high intensity and green representing low intensity for all features. Finally, we obtained 104 clustered images with colored nuclei overlayed with different colors on the glom image according to the assigned cluster label.

We also need to overlay these 104 result images to their original positions on the original kidney image. We load the original image and its corresponding annotation file, and overlap with the clustered image onto the original image based on the position of each annotation. Finally, we save the generated image with all segmented glomerulus regions covered. The pipeline is constructed using scikit-image(van der Walt, 2014).



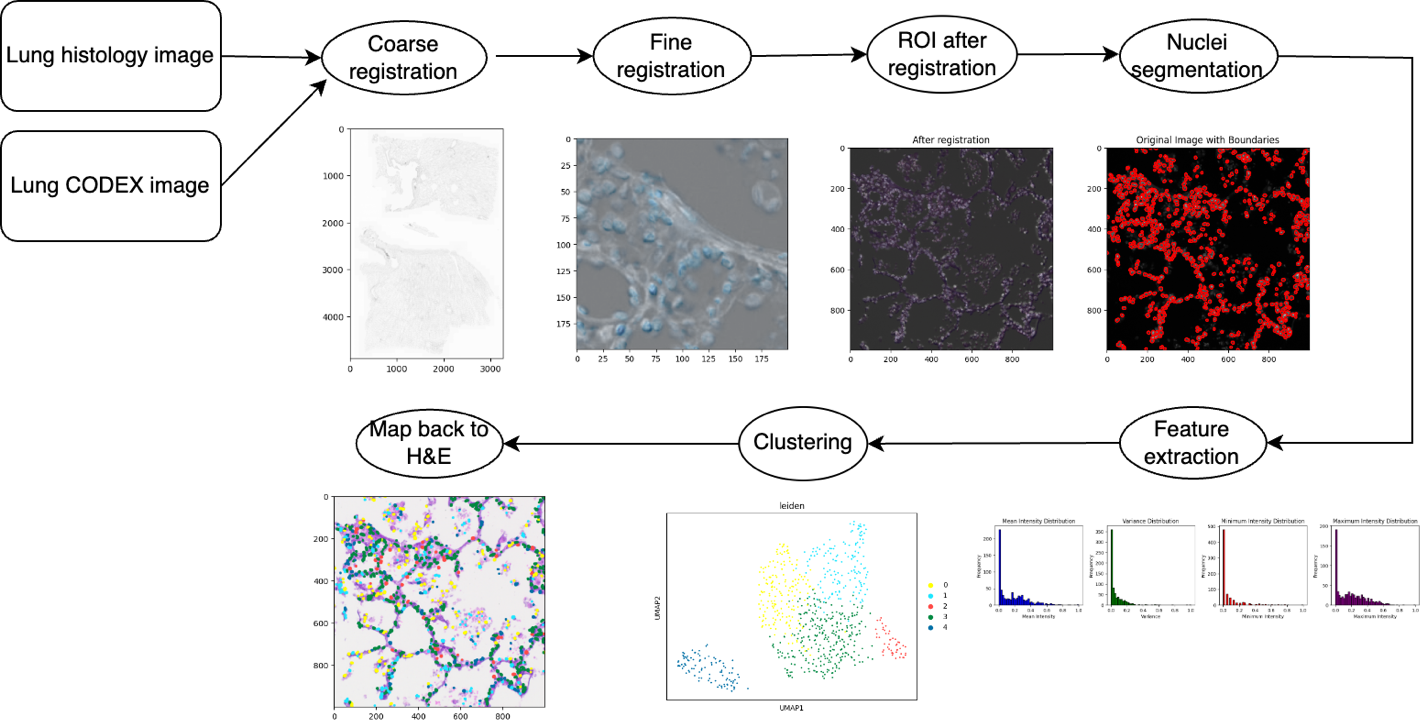
**Fig. 1** | The overview of the Glomeruli pipeline.

* Lung histology and CODEX pipeline (Fig. 2)

Since we will eventually map clustered cells from CODEX onto the histology image, it is necessary to register the two images to ensure that the coordinates are shared across both images. We conduct color deconvolution on the histology image and select the H channel to be registered with the DAPI channel from the CODEX image, as both channels effectively highlight the nuclei. Registration consists of two steps: coarse registration and fine registration. For coarse registration, we pad the DAPI image to match the dimensions of the histology image. Considering the large size of the images and for efficiency reasons, we scale down both images to 1/10 of their original size. Subsequently, we employ phase cross-correlation to determine the shift between the downscaled H channel and the downscaled DAPI image. Using the obtained shift, we construct an affine transformation matrix, which is then rescaled back to fit the original size. The matrix will be applied to the original H&E image to obtain the coarse registered histology image. Based on the coarse registration, we extract a small patch from both the coarse registered histology and DAPI images. We then apply the previous registration method again, resulting in a new affine transformation matrix. This matrix is then combined with the coarse affine transformation matrix, yielding the final transformation matrix. This final matrix is applied to the original histology image for the registered histology image.

Since the histology image has been registered to the DAPI image, we only need to do nuclei segmentation based on the DAPI image. We cropped a small image and used tophat transformation to enhance bright areas smaller than the structural elements (13x13 squares). Then, we apply the Otsu threshold method to the top-hat transformed image to obtain a binary image. Next, we perform morphological processes. A disc-shaped structural element with a radius of 3 pixels is used for binary opening of the binary image obtained. Also, we add binary dilation operation after opening by using a 5x5 square structuring element to enlarge the nuclei, approximating the cell boundaries. Finally, we compute the distance transform on the binary mask and detect local maxima to apply watershed segmentation to separate the touching nuclei.

Next step, we use the same feature extraction as the kidney pipeline. We utilize UMAP(Leland McInnes 2020) for dimensionality reduction and Leiden(Traag et al., 2019) clustering to derive cluster labels, which are then assigned to each nuclei region. The feature matrix still consists of the average intensity, variance, minimum intensity, and maximum intensity, but comes from the 35 channels of the CODEX image. Finally, we map the different cluster labels back to the histology image, assigning distinct colors to each cluster for visualization. The pipeline is constructed using scikit-image(van der Walt, 2014), OpenCV(Bradski, 2000) and Scanpy(Wolf et al., 2018).



**Fig. 2** | The overview of the Lung pipeline.

**Conclusion**

From the two pipelines’ results, the pipelines we created have successfully completed the image processing and nuclei clustering tasks of the input images. They can be used to help researchers find the distribution of different types of cells more conveniently and clearly.

**Contribution**

Xin Ma: H&E preprocessing, registration of lung’s images, Leiden clustering, hyperparameter tuning, map back code, lung pipeline optimization

Yuyang Zhang: DAPI preprocessing, nuclei segmentation, feature extraction, K-means clustering, glom annotation, kidney with glom pipeline optimization

**Reference**

Bradski, G. (2000). opencv\_library. In: Dr. Dobb's Journal of Software Tools.

Leland McInnes , J. H., James Melville. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep, 9(1), 5233. <https://doi.org/10.1038/s41598-019-41695-z>

van der Walt, S. a. S., Johannes L. and Nunez-Iglesias, Juan and Boulogne, François and Warner, Joshua D. and Yager, Neil and Gouillart, Emmanuelle and Yu, Tony and the scikit-image contributors. (2014). scikit-image: image processing in Python. In.

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol, 19(1), 15. <https://doi.org/10.1186/s13059-017-1382-0>