An Analysis on the Classification and Segmentation of Cancerous Cells from Histopathological Images

Melisa Civelekoglu, Gabriel Gidoiu, Merritt Jiang, Luca Monteferrante

I. INTRODUCTION

In recent years, digital whole slide images (WSI) taken of pathological specimens under the microscope, namely histopathological images, have been analyzed based on machine learning (ML) algorithms to assist tasks such as cancer and infection diseases diagnosis and prognosis prediction [1]. This has been specifically useful as, given the large number of slides a pathologist needs to analyze, the pathologists fatigue level as well as variability across different pathologists have caused discrepancies in agreement, potentially leading to misdiagnosis [2].

Our project focuses on building and analyzing a machine learning tool to help with cancer diagnosis by analyzing histopathological images across different areas of the body (lungs, colon, mouth, breasts) (see Appendix A). Unlike previous studies that focus on cancer classification and segmentation of one organ, our goal was to determine which region(s) of the body will benefit most from ML-based classification and segmentation using a deep learning model that can label cancer types accurately and mark regions where cancer cells are located in the tissue. By automating these tasks, the tool can save doctors valuable time by providing useful insights, such as the size, shape, and location of tumors, making the diagnostic process faster and supporting more tailored treatment options for patients based on detailed tumor characteristics.

Our data includes images of different organ tissues which are either normal or have cancerous cells (lung has various types). Next, we apply a semantic segmentation to the images in order to locate the cancerous cells in the tissue. We utilize the Segment Anything Model (SAM) for this purpose, as SAM is a well-trusted segmentation software that works with industry-level accuracy. Finally, we verify which region of the body is best suited for an ML implementation.

Finally, we evaluate and discuss the implications of our findings regarding our classification and the segmentation done by SAM, analyzing how well our classification and segmentation was done per body region's cancer by comparing test accuracies between them.

II. LITERATURE REVIEW

There exist several reviews which discuss the history and details of machine learning algorithms applied to histopathological image analysis [3, 4, 5, 6]. Many studies highlight the main challenges that digital histopathological image analysis faces, which include dealing with large dimensionality and multiple resolutions of WSI [6, 7, 8] and accounting for human errors in histopathological image collection [9] (see Appendix B.1).

Typically, histopathological image analysis using ML includes pre-processing, such as sampling mini patches which help isolate the regions of interest, and feature extraction [1]. One example of a preprocessing transformation used for classifying and/or segmenting histopathological images is color conversion [5, 6] (see Appendix B.2). Another transformation is patch-based

analysis, where researchers feed only patches of the images which they find relevant to the tumor classification into the model [11, 12] (see Appendix B.3).

For classification, papers in this field tend to use: Accuracy, Precision, Sensitivity, F1 Score, Specitivity [6] for evaluating performance. To evaluate segmentation, commonly used metrics are the Dice Score Coefficient, and the Jaccard Index.

In terms of application, researchers generally use segmentation and classification for objectives such as identifying cancers in histopathological images. In general for both parts, every ML concept has been applied. Whether that be unsupervised learning [16], weakly/semi supervised learning [15, 17], reinforced learning [19, 20], and supervised learning. In supervised learning, transferred learning [21, 22], convolutional neural network learning [13, 14], graph-based learning [24, 25], adversarial learning [26], and recurrent learning [23]. There is also research done that combines multiple topics in hybrid models [6]. For segmentation tasks, papers that use hybrid and/or ensemble methods seem to gain the highest accuracy [6]. Convolutional networks can have good accuracies as well, but taking concepts from recurrent learning can help results [6]. For classification, dual-stream approaches [18] as well were attempted which were good, but not the best and took a long time to train/use [6]. The highest performing classification models were all supervised learning, specifically papers using graph-based and convolution-based learnings.

Most existing methods focus on single organs or specific tasks, which leaves a gap when it comes to handling multi-organ data effectively. Recently, Ignatov et al. published a paper outlining their model DeepCMorph pre-trained on 32 different cancer types across certain body regions, which aimed to provide a generalized pre-trained model to be used for smaller microscopy datasets [27]. While they trained and evaluated their model on a collective data containing different body regions, our goal is to train and fine-tune the same model for each body region separately, comparing their accuracies to determine how applicable they are for a ML implementation. Furthermore, most works we have discussed segment and later classify their data, whereas we classify before our segmentation task, as we believe that classifying could lead to better results without any prior beliefs, which could therefore give a potentially significant improvement in segmentation performance. Specifically for segmentation, our method differs from prior papers, as we will be using SAM [28] while they are segmenting the images differently.

III. METHODOLOGY

We develop an ML tool aimed at assisting cancer diagnosis, type classification and marking regions of cancerous tissue through analysis of histopathological images collected across body regions, including mouth, colon, breasts and lungs. We build and fine-tune a convolutional neural network (CNN) for classification and utilize the Segment Anything Model (SAM) for segmentation.

A. CLASSIFICATION

For the classification task, we use a binary or multi-class classification approach (depending on which dataset we use). For those with more than 2 labels, multi-class classification allows us to not only classify each input image as having a cancerous tissue or not, but to assign the type of tumor. We use Cross-Entropy as our loss function.

Our CNN architecture consists of multiple convolutional layers, which are each followed by a batch normalization, ReLU activation and max pooling. Batch normalization is done to help prevent overfitting and accelerate the learning process. To help further mitigate the overfitting, we apply dropout at rates ranging from 10% to 30% progressively throughout the network. Max pooling with a pool size of 2×2 and a stride of 2 is applied to reduce the spatial dimensions of feature maps. Additionally, zero-padding is applied to ensure features that exist at the edges, such as tumor boundaries or tissue structures, are preserved and to control the size of the output feature map. Finally, we use the ADAM optimizer.

Our model contains various hyper-parameters which need to be fine-tuned. These include the image size, probabilities for transformations, the batch size, the pooling and stride magnitude, the amount of convolutional layers, the amount of features in each layer, the kernel size, the dropout probabilities, the weight decay in L2 regularization, etc. (see Appendix C.1 for specific values).

As WSI's are very large-sized images, we ran into the issues described by previous works (see Appendix B.1), leading to our program to crash, throwing an error message from the Numpy library, due to high memory usage. To overcome this, we resize the images to smaller sizes, such as 64×64. Note that training on less images with bigger sizes resulted in worse accuracy than more images with smaller sizes, therefore we concluded with resizing. We found that a 5% chance for transformation allowed our model to combat overfitting on the data while still learning from it. We chose a batch size large enough for the model to learn from while being small enough for it to have many loops in one epoch. The parameters inside the hidden layers were selected through trial and error, according to which values allowed better performance specific to each body region, while keeping the general architecture similar to be able to compare between their results accurately.

B. SEGMENTATION

As SAM only takes images of size 256×256, we rescale our reduced images to the relevant size. We follow the recommendation from [6] to change the color space of the images when doing segmentation and apply color conversion to the images from RGB to LUV.

We use the built-in automatic mask generator function of SAM, as images from our datasets do not come with ground truths for each pixel. We adjust key parameters such as predicted intersection over union and stability score thresholds (see Appendix C.2 for final values), in order to allow the masks to become much more reflective of what a human may consider to be the ground truth (as we cannot determine the actual accuracy of our generated masks due to lack of pixel-wise ground truths). Though there was a visualization issue, we fixed it by implementing the relevant code snippet from [29], which is used only for displaying masks and not for generation.

Note that we experimented with many methods for segmenting our images. Unfortunately, as our publicly available datasets lack pixel-wise ground truths, our classification of the entire image was not useful before segmentation. We originally wanted to train our segmentation model using the predict function, which works similar to how we trained our classification model. However, in order to achieve this, we needed ground truth masks in order for SAM to learn to distinguish between cancerous and benign tissues, which was not possible to acquire.

IV. EVALUATION

To evaluate the performance of our classification task, we include accuracy, precision, sensitivity, F1-score and specificity as our main metrics to compare between different body regions. The following table summarizes these metrics across body regions analyzed in our paper:

Body Region	*Accuracy*	*Precision*	*Sensitivity*	*F1-Score*	*Specificity*
Breast	0.96990	0.82608	0.97080	0.89261	0.96976
Colon	0.96700	0.99681	0.93700	0.96597	0.99700
Lung	0.96613	0.96835	0.96613	0.96606	0.98306
Mouth	0.8730	0.96470	0.86320	0.91110	0.90320

Table 1 - Evaluation Metrics Summary for Cancer in Different Body Regions

For more in-depth analysis and visual interpretations, i.e. plots showing training and validation loss curves, accuracy trends, and confusion matrices, see Appendix D.

As mentioned before, as our datasets do not include pixel-wise ground truths, it is not possible to apply segmentation metrics to our SAM implementation and verify its performance. However, for our sample images with segmentation, see Appendix E.

V. DISCUSSION

Our goal for this project was to determine which among the different body regions we focused on were most fit for an ML implementation. With an accuracy of consistently above 85% for all body regions, our classification model is capable of accurately identifying key features for classification in most cases. The model achieves the highest sensitivity on breast cancer data, which indicates that the model best identifies true positive cases in this region. On the other hand, the model achieves the highest precision and specificity for the colon region, indicating that it is best suited for avoiding false positives in this region. This can be particularly important when the goal is to lower false alarm rates of pathologists misdiagnosing negative colon cancers as positive. Although the lung region has 2 different types of cancerous tissue in addition to the benign tissue, the model shows balanced performance across all metrics in this area, which could indicate that our model is more reliable in both detecting cancerous tissues and avoiding false alarms than other regions we looked at in this paper. For the oral region, the model showed lower performance compared to other regions, likely due to the dataset used for training being smaller. However, it still achieved reasonable accuracy despite this.

Although it is not possible to apply evaluation metrics and verify our segmentation results generated using SAM, our observations based on the alignment of the segmented masks with visible features on the original images can still offer some insight into the model's performance. From our research, we found that the cancerous tissues are ones which include "black dots" in the WSI. According to this, when these black dots are clearly distinguishable from the background, i.e. when they have a high contrast against the image background, SAM performs better in producing more distinctly segmented masks of the relevant areas, in accordance to what a human observer would

consider contains the cancerous tissue. On the other hand, when these "black dots" are less distinct and instead seem to be blending to the background, SAM will produce segmentation masks with less observed accuracy and are not as clear in comparison. This suggests that, without training on images with pixel-wise ground truths, SAM's ability to segment regions may rely on visual contrasts which exist in the input image, thus could limit its applicability to cases where the cancerous regions are much less visually pronounced. Finally, note that due to the 256×256 requirements for image sizes by SAM, we had to re-scale our images across different body regions differently. For example, the breast cancer data had to be upscaled as their sizes were smaller in comparison to colon and lung cancer data which had to be downscaled as their sizes were much larger. This could have also had an impact on the quality of the segmentation mask, as recalling could affect how close or further apart these "black dots" are from each other.

VI. CONCLUSION

Our goal was to develop an ML tool for multi-class classification and segmentation in histopathological cancer images. We used CNN for classification and SAM for segmentation. Overall, we achieved high classification accuracy across all body regions, with good performance in metrics like F1-score, precision, and sensitivity, demonstrating the model's reliability in detecting and classifying cancerous tissues. For segmentation, SAM produced visually meaningful masks when cancerous regions had high contrast (clear "black dots"), but struggled with low-contrast features and larger tumors due to the lack of ground truth masks.

The two main limitations we faced with this project were dataset and time related. Specifically for SAM, our datasets should have included ground truth values at each pixel in order to accurately train and evaluate our implementation. As there are not many relevant publicly available datasets with such ground truth masks, we are unable to draw conclusions about how well separate body regions can be adapted to SAM. However, our segmentation results are still able to highlight an important fact: the undeniable value of pathologists and doctors in providing analysis and annotations of cancer cells in WSI. Even when the goal is to train a strong ML model capable of highly accurate segmentation in order to mitigate misdiagnoses caused by human error [2], the initial annotation and validation from experts in their field is necessary. Furthermore, we were limited by the amount of time we had to complete this project; more time would have allowed us to search for better suited datasets (i.e. contact researchers for datasets they may have access to) and build a further, more tailored analysis using results from a more accurate SAM to analyze correlations between the location and the type of cancer it is, and how the size and shape of the cancer affects our predictions, like we initially hoped to achieve.

Future directions that can be considered include implementing stain normalization to standardize image colors across datasets, integrating recurrent learning with CNN architectures to further enhance feature extraction and classification accuracy and acquiring datasets with pixel-wise ground truth masks which will enable supervised training and robust evaluation of segmentation performance.

Finally, note that our source code link can be found under Appendix F.

REFERENCES

[1] Daisuke Komura, Shumpei Ishikawa, Machine Learning Methods for Histopathological Image Analysis, Computational and Structural Biotechnology Journal, Volume 16, 2018, Pages 34-42, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2018.01.001.

[2] Lin, H., Chen, H., Dou, Q., Wang, L., Qin, J., Heng, P.-A.: ScanNet: A Fast and Dense Scanning Framework for Metastatic Breast Cancer Detection from Whole-Slide Image. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2018)

[3] Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 2017;19:221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442.

[4] Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng 2009;2:147–71. https://doi.org/10.1109/RBME.2009.2034865.

[5] Wu Y, Cheng M, Huang S, Pei Z, Zuo Y, Liu J, Yang K, Zhu Q, Zhang J, Hong H, Zhang D, Huang K, Cheng L, Shao W. Recent Advances of Deep Learning for Computational Histopathology: Principles and Applications. Cancers (Basel). 2022 Feb 25;14(5):1199. doi: 10.3390/cancers14051199. PMID: 35267505; PMCID: PMC8909166.

[6] Abdel-Nabi, H., Ali, M., Awajan, A. et al. A comprehensive review of the deep learning-based tumor analysis approaches in histopathological images: segmentation, classification and multi-learning tasks. Cluster Comput 26, 3145–3185 (2023). https://doi.org/10.1007/s10586-022-03951-2

[7] Takahama, S., Kurose, Y., Mukuta, Y., Abe, H., Fukayama, M.,Yoshizawa, A., Kitagawa, M., Harada, T.: Multi-stage pathological image classification using semantic segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (2019)

[8] McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B., Kovacevic, J.: Automated histology analysis: opportunities for signal processing. IEEE Signal Process Mag. 32, 78–87 (2015). https://doi.org/10.1109/msp.2014.2346443

[9] Raab, S.S., Grzybicki, D.M., Janosky, J.E., Zarbo, R.J., Meier, F.A., Jensen, C., Geyer, S.J.: Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. Cancer 104, 2205–2213 (2005). https://doi.org/10.1002/cncr.21431

[10] Yang, L., Meer, P., Foran, D.J.: Unsupervised segmentation based on robust estimation and color active contour models. IEEE Trans. Inf Technol. Biomed. 9, 475–486 (2005). https://doi.org/10.1109/titb.2005.847515

[11] Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., Gonzalez, F.: High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: application to invasive breast cancer detection. PLoS One. 13, e0196828 (2018). https://doi.org/10.1371/journal.pone.0196828

[12] Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D.E.: Cluster-to-conquer: a framework for end-to-end multi-instance learning for whole slide image classification. arXiv (2021). https://doi.org/10.48550/arXiv.2103.10626

[13] Gu, F., Burlutskiy, N., Andersson, M., Wile´n, L.K.: Multi-resolution networks for semantic segmentation in whole slide images. Computational pathology and ophthalmic medical image analysis, pp
[14] Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J., Shapiro, L.: Learning to Segment Breast Biopsy Whole Slide Images. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2018)
[15] Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W.: CAMEL: A weakly supervised learning framework for histopathology image segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (2019)
[16] Tellez, D., van der Laak, J., Ciompi, F.: Gigapixel whole-slide image classification using unsupervised image compression and contrastive training. Med. Imag. Deep Learn. (2018)
[17] Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.-A.: Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern. 50, 3950–3962 (2020).https://doi.org/10.1109/ tcyb.2019.2935141
[18] Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self supervised contrastive learning. In: Proceedings of the IEEE/ CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
[19] Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Reinforced auto-zoom net: towards accurate and fast breast cancer segmentation in whole-slide images. Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 317–325. Springer International Publishing, Cham (2018)
[20] Xu, B., Liu, J., Hou, X., Liu, B., Garibaldi, J., Ellis, I.O., Green, A., Shen, L., Qiu, G.: Look, investigate, and classify: a deep hybrid attention method for breast cancer classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 914–918. IEEE (2019)
[21] Pham, H.H.N., Futakuchi, M., Bychkov, A., Furukawa, T., Kuroda, K., Fukuoka, J.: Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. Am. J. Pathol. 189, 2428–2439 (2019). https://doi.org/10.1016/j.ajpath.2019.08.014
[22] Wang, P., Li, P., Li, Y., Wang, J., Xu, J.: Histopathological image classification based on cross-domain deep transferred feature fusion. Biomed. Signal Process Control 68, 102705 (2021). https://doi.org/10.1016/j.bspc.2021.102705
[23] BenTaieb, A., Hamarneh, G.: Predicting Cancer with a Recurrent Visual Attention Model for Histopathology Images. In: Medical Image Computing and Computer Assisted Intervention Cluster Computing (2023) 26:3145–3185 3179 123 MICCAI 2018. pp. 129–137. Springer International Publishing (2018)
[24]Agarwal, N., Balasubramanian, V.N., Jawahar, C.: v: Improving multiclass classification by deep networks using DAGSVM and Triplet Loss. Pattern Recognit. Lett. 112, 184–190 (2018). https://doi.org/10.1016/j.patrec.2018.06.034
[25] Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L.: A survey on graph-based deep learning for computational histopathology. Comput. Med. Imaging Graph. 95, 102027 (2022). https://doi.org/10.1016/j.compmedimag.2021.102027

[26] . Xue, Y., Ye, J., Zhou, Q., Long, L.R., Antani, S., Xue, Z., Cornwell, C., Zaino, R., Cheng, K.C., Huang, X.: Selective synthetic augmentation with HistoGAN for improved histopathology image classification. Med. Image Anal. 67, 101816 (2021). https://doi.org/10.1016/j.media.2020.101816
[27] Ignatov, A., Yates, J., & Boeva, V. (2024). Histopathological Image Classification with Cell Morphology Aware Deep Neural Networks. arXiv. https://arxiv.org/abs/2407.08625
[28] Ma, J., He, Y., Li, F. et al. Segment anything in medical images. Nat Commun 15, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
[29] Fatauil, H. (2023, August 10). *SAM (Segment Anything Model)*. Medium. https://medium.com/@hasfatauil12/sam-segment-anything-model-d4f541165f6b

APPENDIX

A DATASETS

Note that currently we are using datasets from the following resources:

- *Breast Cancer Dataset*. Kaggle. https://www.kaggle.com/datasets/akhilbs/ breastcancer
- *Histopathologic Oral Cancer Detection using CNNs*. Kaggle. https://www.kaggle.com/datasets/ashenafifasilkebede/dataset
- *Lung and Colon Cancer Histopathological Images.* Kaggle. https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

B ADDITIONAL LITERATURE REVIEW

The following are sections from our literature review:

B.1 MAIN CHALLENGES

Many studies highlight the main challenges that digital histopathological image analysis faces. Firstly, the WSI have very large dimensionalities since they contain multiple resolutions, and even with GPU-based models that can gain up to 40 times acceleration over CPU-based models, WSI analysis still faces a large computational cost problem [6]. Furthermore, WSI at highest resolutions are far too big to be compatible with certain deep learning models, and feeding lower-resolutions to the model can result in cellular level features to be lost, leading to lower performance [7, 8]. Additionally, specifically for WSI which look at cancer cells, the tumor usually occupies only a small fragment of the image, which can lead to data imbalance when it is compared to the background, foreground, and normal tissue areas [6]. Finally, image acquisition is a manual process prone to human error and can alter the appearance of the tissue by introducing anomalies on the WSI [9].

B.2 COLOR CONVERSION

Similar papers have used stained images which initially have RGB inputs and convert them into LUV space [10]. Color conversion allows different parts of the image to have a more pronounced contrast than in the original image thus ultimately after parametrizing the gradient flow estimation will better differentiate the concerned part (e.g. the tumor) from the environment (e.g. the other tissues surrounding it).

B.3 MINI-PATCHING

As mentioned before, high resolution WSI's are computationally demanding and can include a lot more information than what's necessary for tumor classification. To resolve this issue and make the models more efficient, researchers have fed only patches of the images that they find relevant to the tumor classification into the model [11, 12]. This is done with a variety of sampling methods ranging from simpler ones such as regular or uniform to more complex ones that are

customized strategies. These previous transformations are the most common transformations specifically used for histopathological images, but standard transformations such as rotating, flipping, gaussian noise, etc. are all valid transformations which can be used.

C METHODOLOGY: FINE-TUNED PARAMETERS ACROSS DATASETS

The following parameters were optimized in order to achieve highest possible accuracy for classification and segmentation tasks for each body region.

C.1 FINE-TUNED CLASSIFICATION

Body Region	*Batch Size*	*Epochs*	*Image Size*	*Convolution Layers*	*Dropouts*	*Learning Rate*
Breast	128	25	50 × 50	5	0.4, 0.3, 0.25, 0.15	0.001
Colon	128	25	64 × 64	5	0.4, 0.3, 0.25, 0.15	0.001
Lung	64	50	64 × 64	5	0.3, 0.3, 0.2, 0.15, 0.15, 0.1	0.01
Mouth	128	25	50 × 50	3	0.5	0.001

Table 2 - Fine Tuned Parameters for CNN Model

C.2 FINE-TUNED SEGMENTATION

Body Region	Predicted Intersection over Union Threshold	Stability Score Threshold
Breast	0.9	0.95
Colon	0.855	0.905
Lung	0.9	0.9
Mouth	0.95	0.925

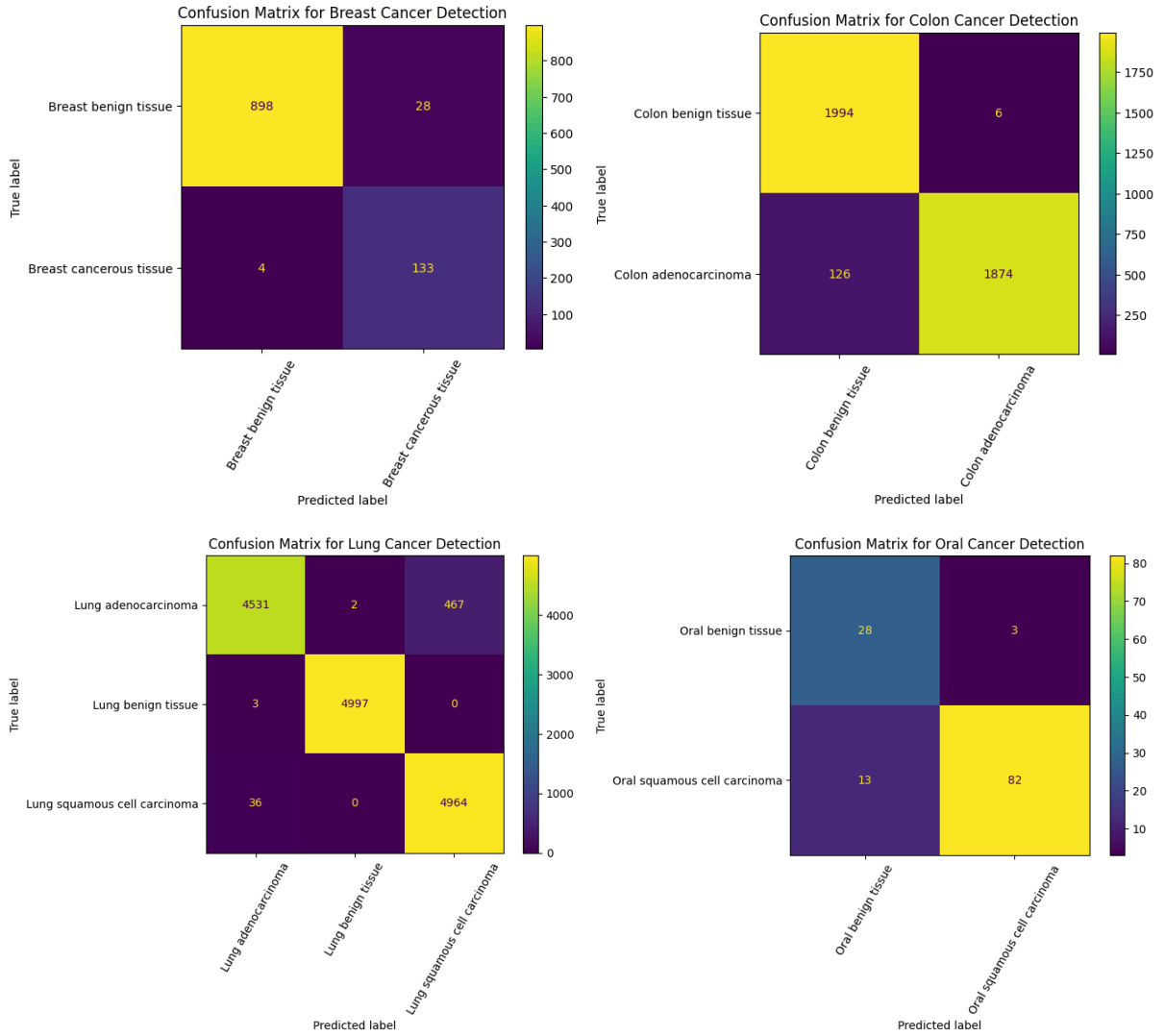
Table 3 - This table demonstrates the final fine-tuned hyperparameters of the models.

D EVALUATION: MORE METRICS

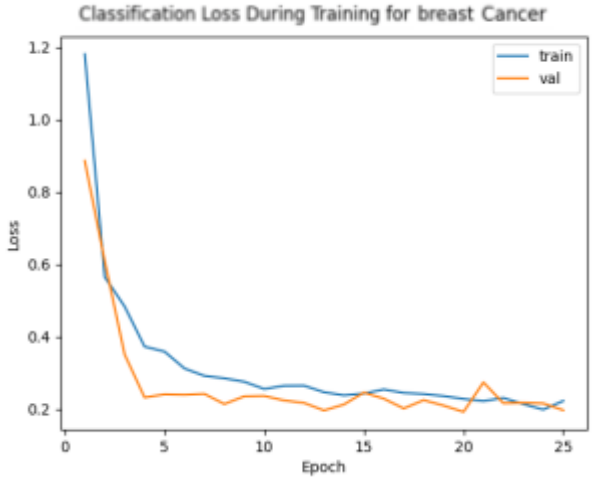
To summarize Table 1: All the regions besides the mouth/oral gave similar accuracies. The mouth gave a much lower accuracy due to the lower amount of images used in training. For precision, this metric is used to compare the cancerous images we classified correctly compared to all the cancerous ones we classified . The colon model was best suited for this task while the breast was the worst. This low amount is most likely due to the up-scaling of the images. Sensitivity is used to compare the cancerous images we classified correctly compared to all the cancerous images in the

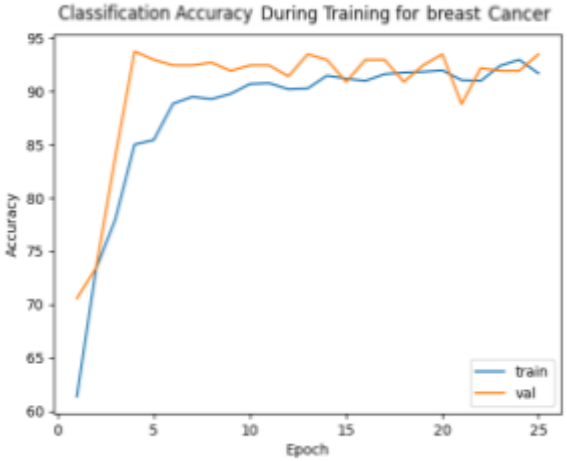
dataset. The breast dataset excelled at this while the mouth did the worst. The lung dataset excelled in the F1 score while the breast did the worst in that. F1 score aims to compare the cancerous images we classified to all the images that we incorrectly classified. Finally the last classification metric was the specificity which the colon model did the best and the mouth model did the worst. Specificity compares the non-cancerous images we classified to all non-cancerous images. It was always the breast and mouth models which did the worst, which can be attributed to the originally small image sizes in the breast dataset and the lack of images to train on in the mouth dataset.

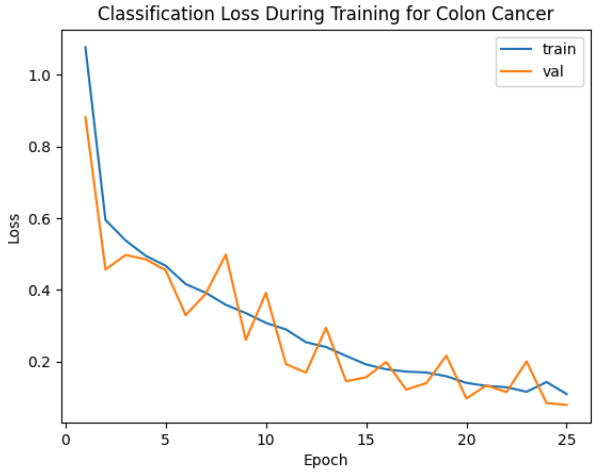
The following confusion matrices add a depth of analysis, which would allow us to identify areas of misclassification, such as whether certain tumor types are more prone to errors due to overlapping features or insufficient data representation.

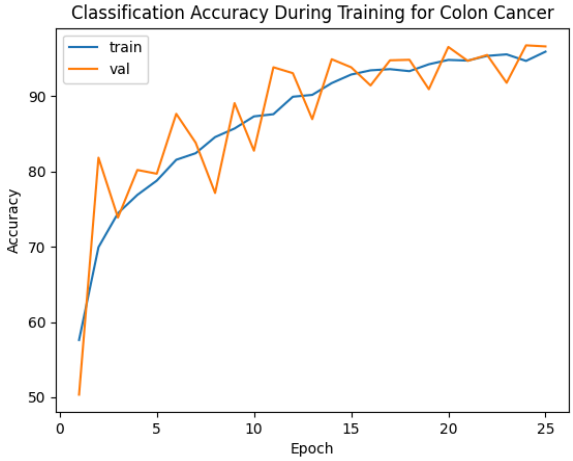


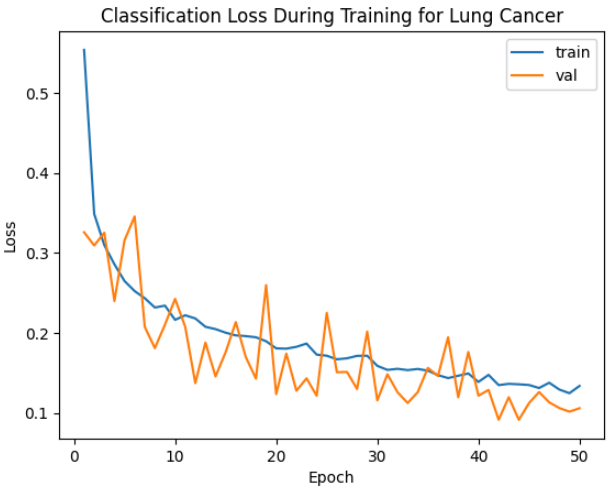
Additionally, the following plots show the training and validation loss and accuracy curves, which helped us in fine-tuning and gave insights into the performance of our model on each dataset.

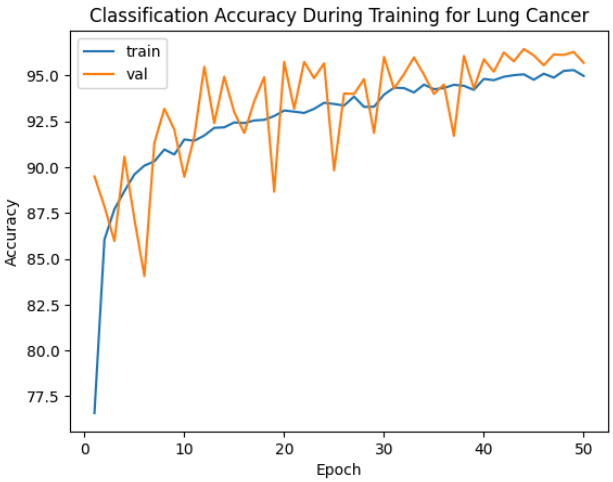

Classification Loss During Training for breast Cancer
train
val
Loss
Epoch

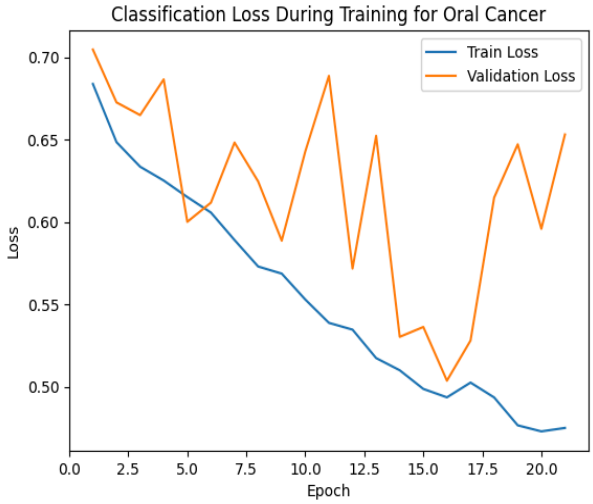

Classification Accuracy During Training for breast Cancer
train
val
Accuracy
Epoch

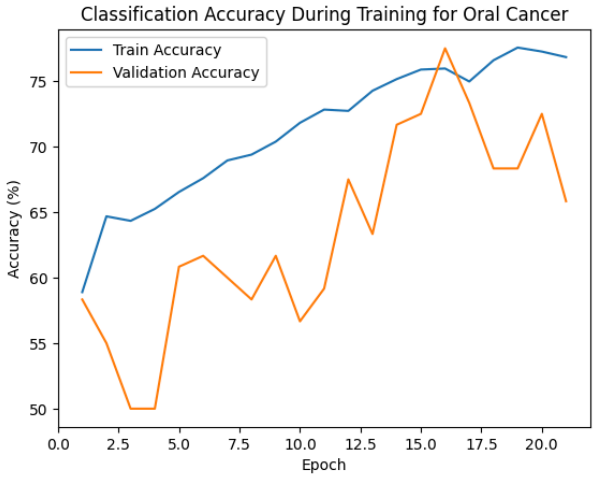

Classification Loss During Training for Colon Cancer
train
val
Loss
Epoch


Classification Accuracy During Training for Colon Cancer
train
val
Accuracy
Epoch


Classification Loss During Training for Lung Cancer
train
val
Loss
Epoch


Classification Accuracy During Training for Lung Cancer
train
val
Accuracy
Epoch



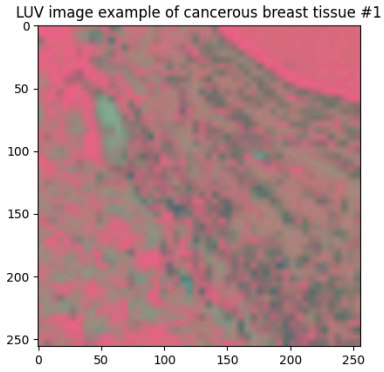


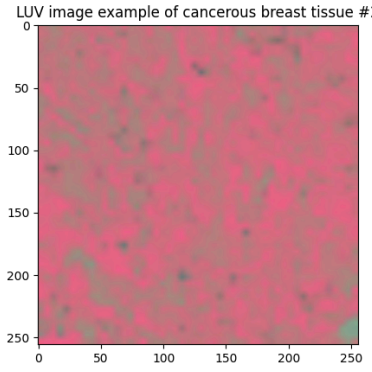
E SEGMENT ANYTHING MODEL RESULTS

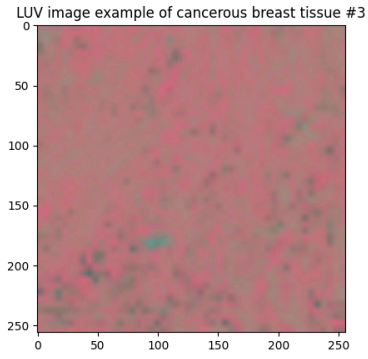
The following are our resulting segmentation masks generated by SAM against the original, color converted images, as well as some short descriptions of observations made for each dataset.

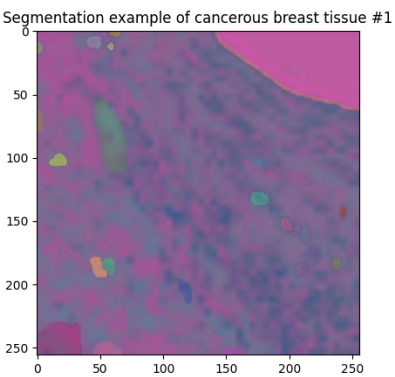
E.1 BREAST DATASET

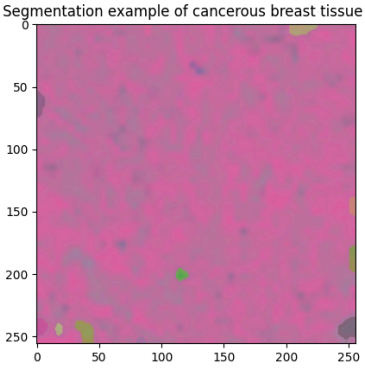
This dataset had small images, causing us to need to upscale the size for it to work with SAM. Thus the segmentations do not accurately describe the image well. That much can be seen even without metrics. Larger images would definitely help the segmentation as qualities such as distance between cancerous cells and color gradients seem promising.

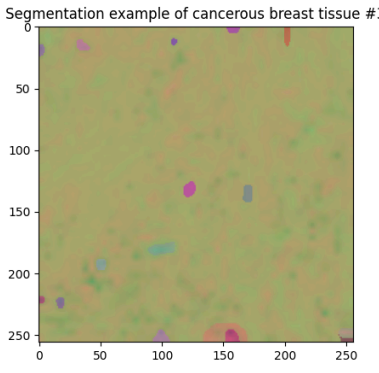


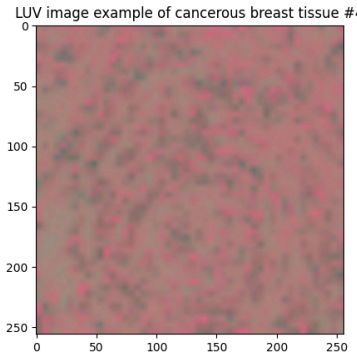


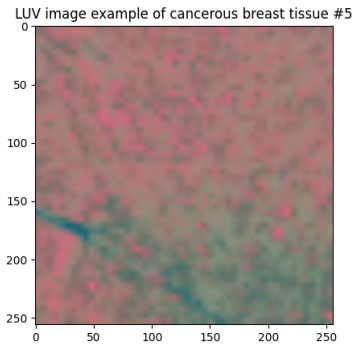


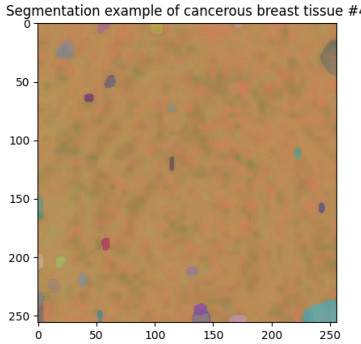


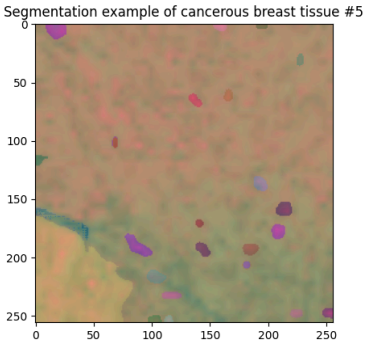






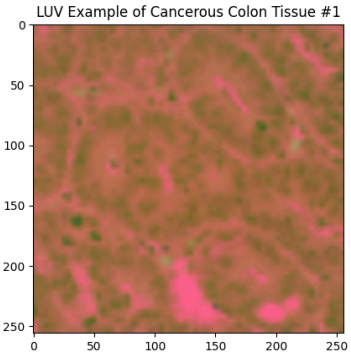


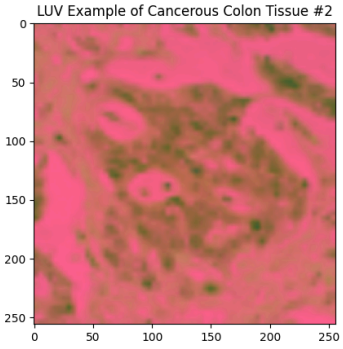


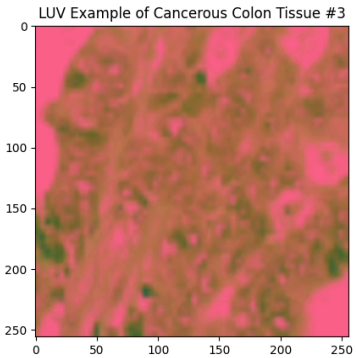


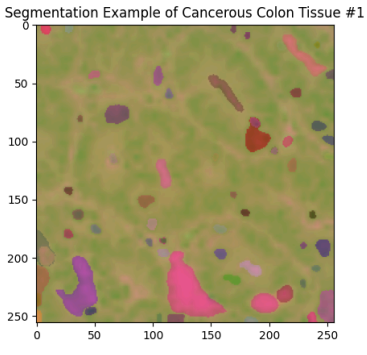
E.2 COLON DATASET

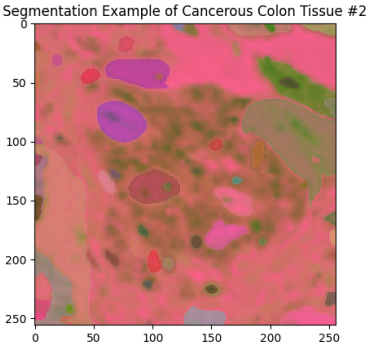
The colon dataset had some promising segmentations on images that had the cancerous tissue more spread out. But there were also some that did not segment it very well. Without having an evaluation model, it's hard to know how good the segmentation is. Visually some look very good and almost perfect, while some look terrible. We'd need the ground truths per-pixel to know.

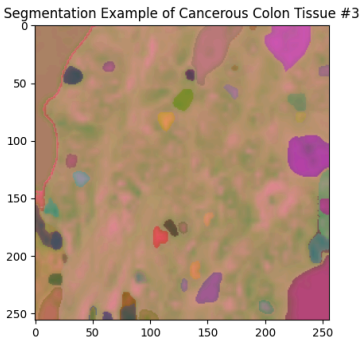


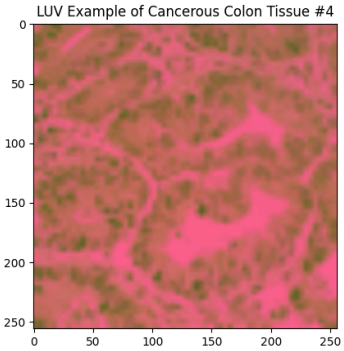


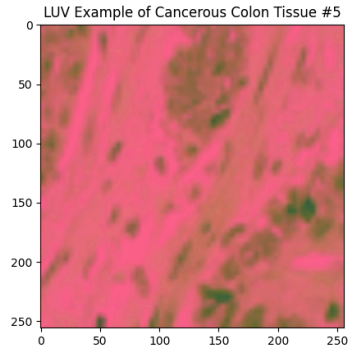


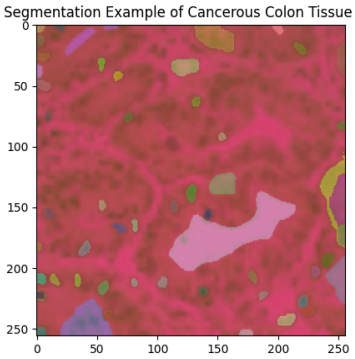


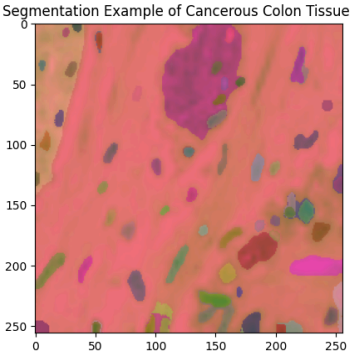






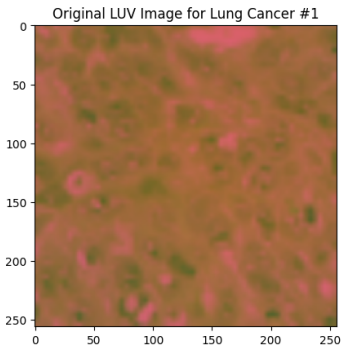


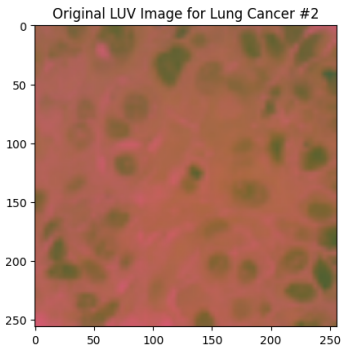


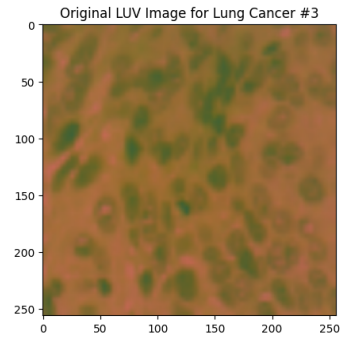


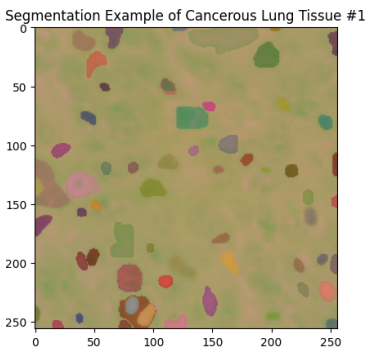
E.3 LUNG DATASET

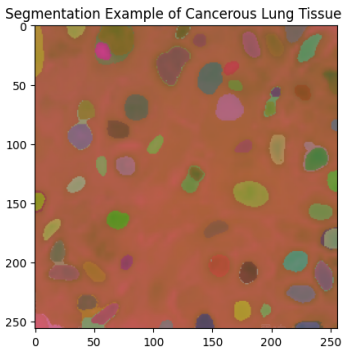
For lung segmentation, our observations suggest that larger, or rather more zoomed-in, tumors are harder to segment using SAM, while smaller tumors have a higher likelihood of being segmented properly. Additionally, some images have tumors with slightly different colors, since this dataset includes multiple tumor types (lung adenocarcinoma and squamous cell carcinoma) which likely affect how it is perceived by SAM.

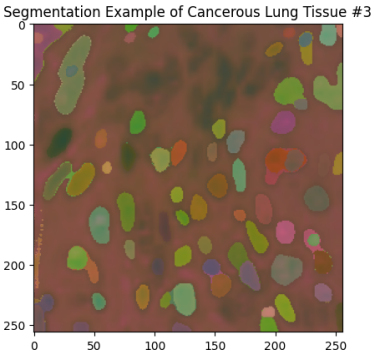


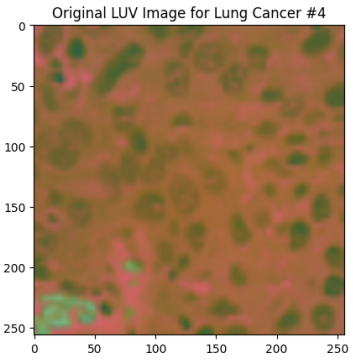


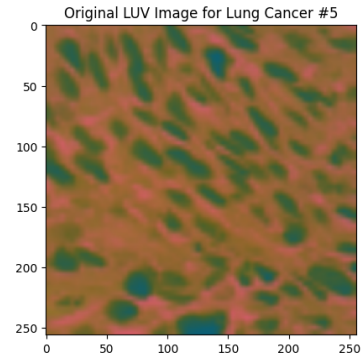


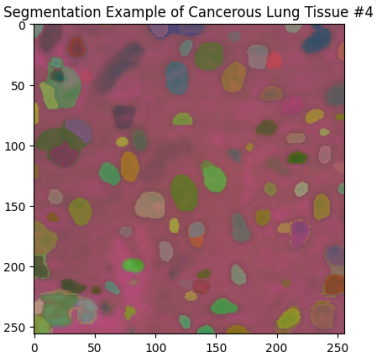


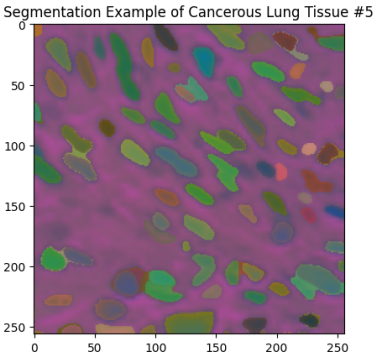






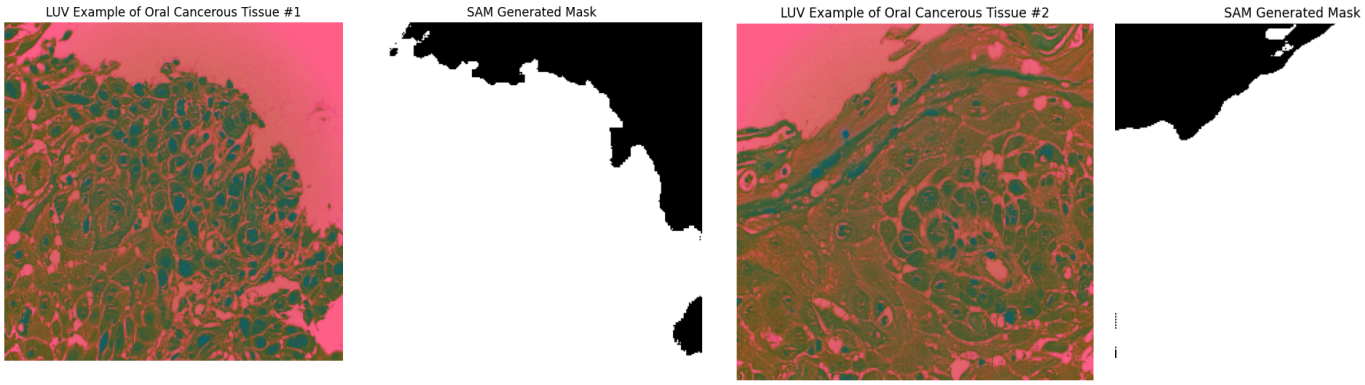


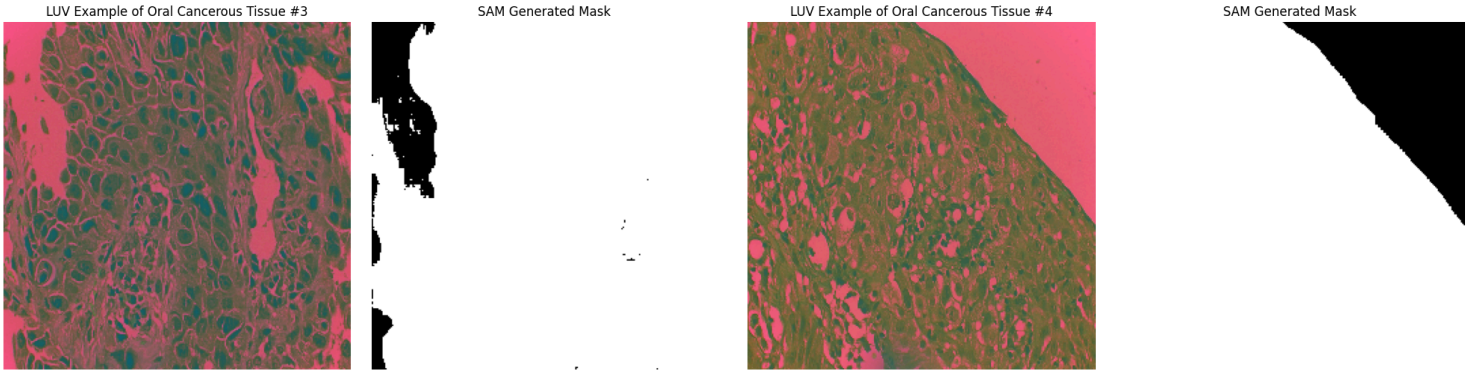


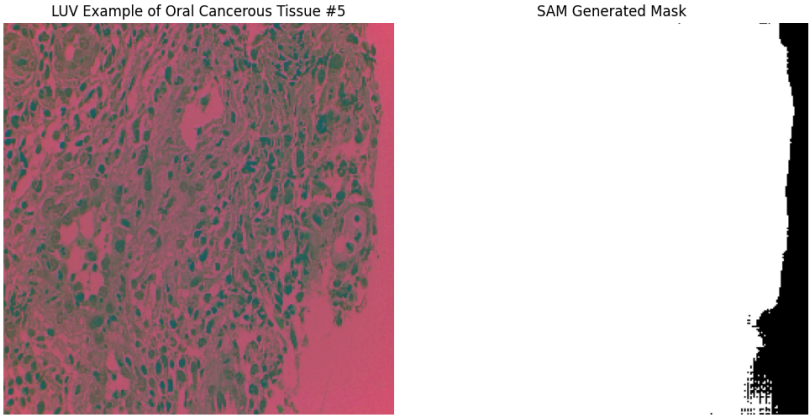


E.4 ORAL DATASET

For the oral dataset, the SAM-generated mask primarily separates the background but fails to capture the dense cellular structures of the tissue. This is likely due to resizing the high-resolution images to 256x256, causing a loss of fine details. The complex patterns and subtle color variations in oral tissue further challenge accurate segmentation.







F LINK TO SOURCE CODE

Our code includes separate files for classification and segmentation of every body region, and can be accessed from the following link:

https://github.com/MeRriTtJL/Comp451-Final-project/blob/main/