InstaDeep Take Home Test
SPADESegResNet: Combining power of SPADE and ResNet for Breast Cancer Semantic Segmentation

Srijay Deshpande

1. PROBLEM STATEMENT

The take-home test involves developing a multi-label classifier to assign each segmentation label a corresponding category. In other words, the task is about classifying each pixel within histology images into distinct tissue regions or a semantic segmentation of breast cancer histology images.

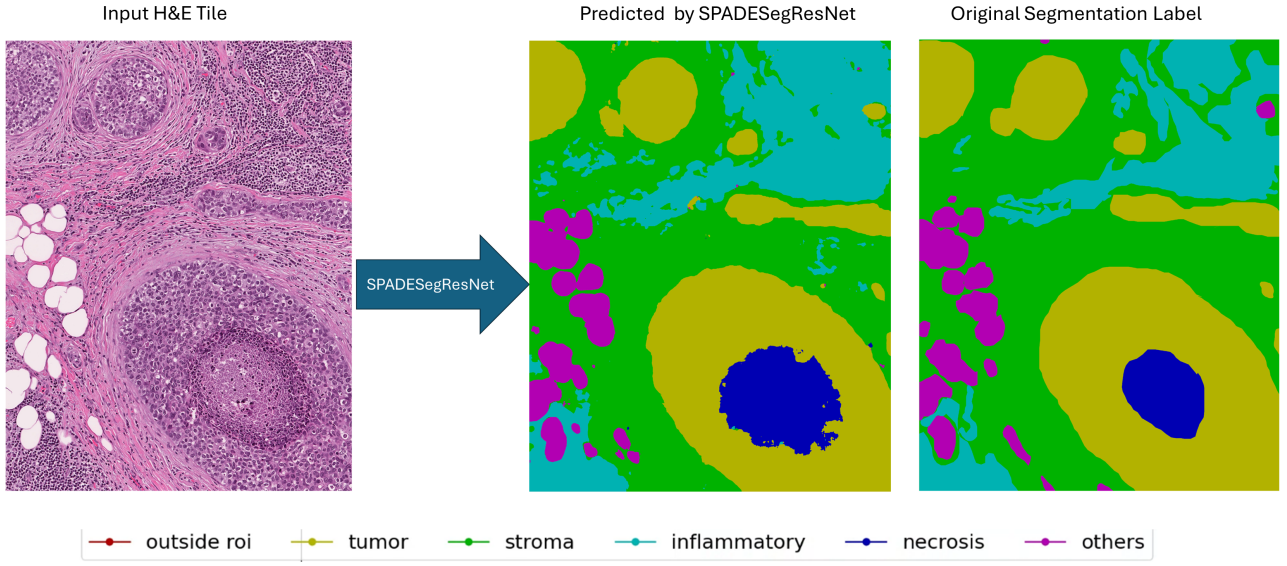

Figure 1: The proposed SPADESegResNet model can be used for semantic segmentation of breast cancer histology images

2. ABSTRACT

Annotating tissue regions within whole-slide histology images poses a significant challenge for clinical experts and practitioners. In this assignment, I propose the SPADESegResNet model for classifying image pixels into distinct tissue regions. This model effectively segments tissue regions while preserving finer details. The framework is assessed both visually and quantitatively using performance metrics such as the Dice score, AUC-ROC, and pixel classification accuracy. The model achieves a Dice score of 0.77 and an AUC-ROC of 0.79 for the identification of tumor regions. Furthermore, the model demonstrates moderately superior performance compared to baseline models, including UNet and its enhanced version, UNet++. The implementation of the proposed SPADESegResNet framework is accessible at `https://github.com/Srijay/SPADESegResNet`.

3. RELATED WORK

Accurate semantic segmentation of tissue regions inside the whole slide images or exhaustive pixel level classification is a challenging task in computational pathology. Several models have been developed in the computer vision community for semantic segmentation tasks. For instance, Ronneberger et al. proposed the UNet.[1] UNET represents a U-shaped encoder-decoder network architecture, comprising encoder and decoder blocks interconnected through a bridge. The encoder network reduces spatial dimensions by half and doubles the filter count at each

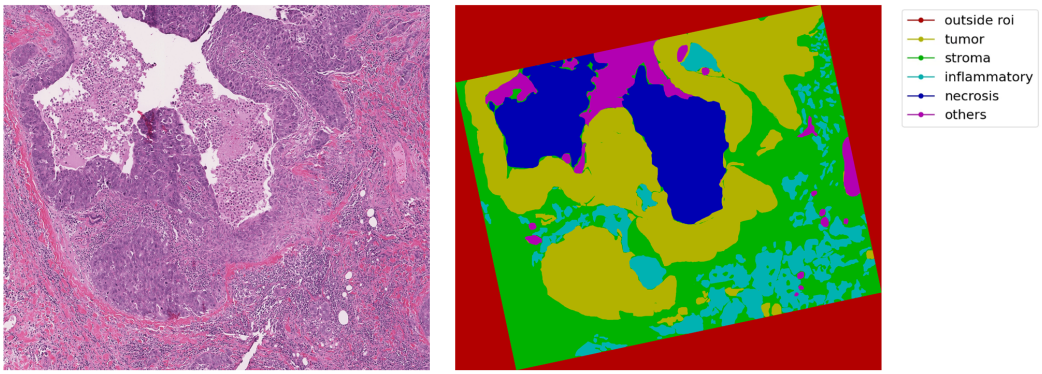

Figure 2: Sample of breast cancer tissue ROI and corresponding segmentation labels for distinct tissue regions or tissue types from the BCSS dataset

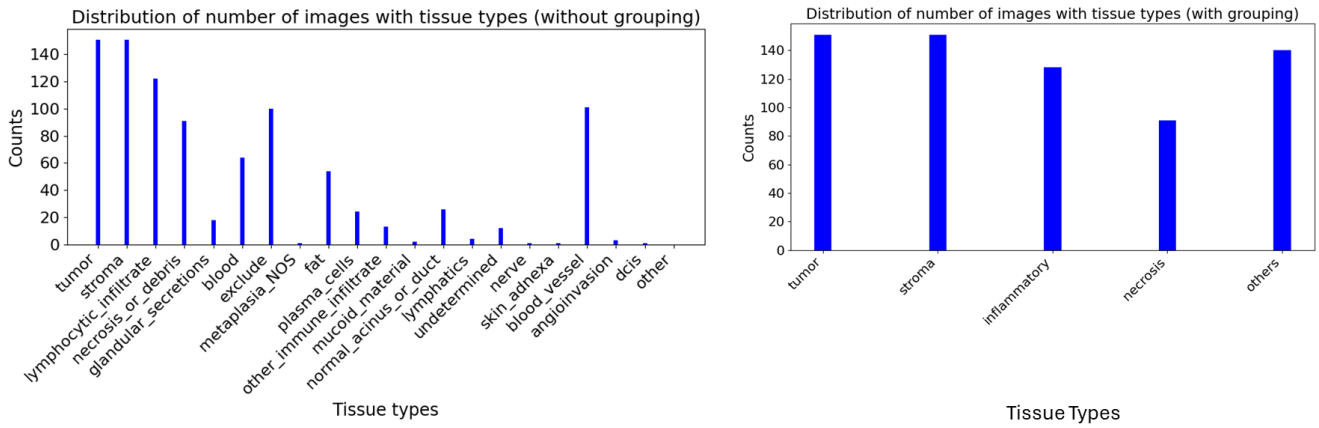

Figure 3: Bar graph quantifying the number of images containing specific tissue regions or types across the whole BCSS dataset

encoder block. Conversely, the decoder network doubles the spatial dimensions and halves the number of feature channels. UNet has demonstrated enhancements over fully convolutional networks (FCN)..[2]

In 2019, Zhou et al. developed the UNet++ or nested UNet,[3] a semantic segmentation model based on the UNet. It is a deeply-supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested skip pathways instead of direct skip connections used in the UNet. To deal with the annotations scarcity for tissue regions, Mostafa et al. proposed an efficient interactive segmentation network called EfficientUNet[4] that requires minimum input from the user to accurately annotate different tissue types in the histology image.

While the above networks performed well for semantic segmentation, I felt an absence of direct visibility of the input tissue image in the deeper layers of the deep neural frameworks, which could potentially enhance segmentation performance. The normalization layers in models tend to wash way some key details from the input image till the information reaches the final layer. Therefore, in this assignment, I attempt to develop a model called SPADESegResNet for the task of breast cancer tissue segmentation. The proposed model is inspired from both SPADE[5] and ResNet[6] networks.

4. DATASET ANALYSIS & PREPROCESSING

The Breast Cancer Semantic Segmentation (BCSS) dataset[7*] used in this task contains over 20,000 annotated regions of interest (ROI) from 151 Hematoxylin and Eosin (H&E) stained whole slide images (WSI) with the

*https://bcsegmentation.grand-challenge.org/

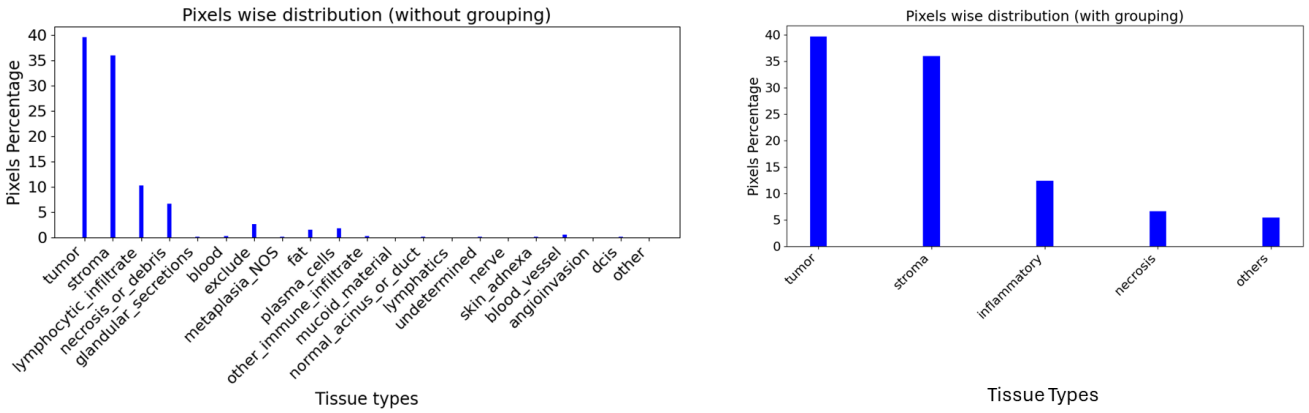

Figure 4: Bar graph showing the percentage of pixels corresponding to distinct tissue types across the whole BCSS dataset

same number of patients from the TCGA[8] dataset. This dataset was annotated through the collaborative effort of pathologists, pathology residents, and medical students using the Digital Slide Archive.[7]

The ROIs within the dataset have different sizes, ranging from a maximum image size of $9248 \times 9335 \times 3$ pixels to a minimum image size of $957 \times 2359 \times 3$ pixels, all captured at a resolution of 0.25 microns per pixel (MPP). The dataset predominantly consists of 138 Whole Slide Images (WSI) acquired at a magnification of $40\times$, with only 13 WSIs obtained at a magnification of $20\times$. Each image in the dataset has a corresponding segmentation map or segmentation mask, where individual labels represent 21 distinct tissue types. The sample image-annotation pair can be seen in Figure 2.

To assess the distribution of tissue regions across the images in the BCSS dataset, the analysis involves quantifying the number of images containing specific tissue regions or types. This data is then visually represented in a bar graph, as depicted in Figure 3 (left). The bar graph clearly indicates that certain tissue types, such as dcis, angioinvasion, lymphatics, and nerves, are infrequently observed in the images, and they appear to be heavily underrepresented. In accordance with the methodology outlined in the original dataset paper,[7] a grouping strategy is employed for tissue types. Regions associated with rare classes are grouped with predominant classes where applicable. Specifically, angioinvasion and dcis are grouped with tumor, while lymphocytes, plasma cells, and other immune infiltrates are grouped with inflammatory infiltrates. The resulting bar graph for the grouped classes is presented in Figure 3 (right).

Despite the somewhat balance in the distribution of the number of images containing specific tissue types achieved through grouping, it is equally important to know the occurrence distribution of these tissue types in terms of pixels inside images. This analysis provides a more comprehensive understanding of how these tissue regions are distributed across the images in the dataset. Consequently, we present plots illustrating the percentage of pixels corresponding to distinct tissue types across the whole dataset with and without grouping tissue types. These bar graphs are depicted in Figure 4. It can be easily inferred from these histograms that data is heavily imbalanced towards tissue types necrosis, inflammatory, and others. Another potential issue with a few examples from the BCSS dataset is with the labeling of the 'don't care' or outside region of interest (outside ROI) segmented region. In some cases, this region is labeled from H&E tissue samples where tissue is present. For instance, consider the example shown in Figure 2, where a substantial amount of tissue region is mapped to the outside ROI region, as highlighted in red.

For experimental purposes, the ROIs and their corresponding segmentation labels are divided into distinct training and testing sets, comprising 105 images for training and the remainder 46 images for testing purposes. Special attention has been given to ensuring a proportionate representation of all five classes in both sets. Subsequently, 5343 overlapping tiles are extracted from the training set, referred to as training images throughout this report. Each tile measures 768×768 pixels. Similarly, 2516 tiles of the same size are extracted from the testing set for testing purposes, referred to as testing images. Images with over 90% of the area of the "don't

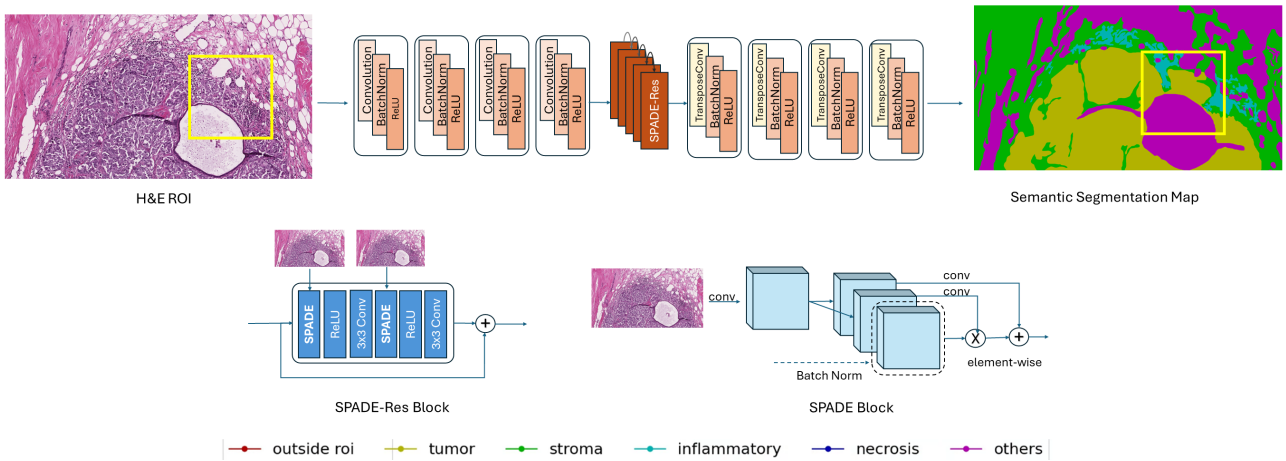

Figure 5: An overview of the proposed SPADESegResNet framework for computing semantic segmentation labels from input H&E regions of interests (ROI). The initial step involves passing the 768×768 pixels tile, shown in yellow-bordered region, through a sequence of convolutional layers followed by batch normalization and ReLU activation. Subsequently, the resultant deep representation undergoes processing through SPADE-Res blocks, incorporating SPADE layers in conjunction with convolutional layers. In the SPADE layer, the input image undergoes convolution operations to generate two modulation parameters, which are element-wise multiplied and added to the normalized activation respectively. The output representation from SPADE-Res units is further processed through a series of transpose convolution layers, batch normalization, and ReLU activation. This series of operations collectively upsample the image, resulting in the generation of the semantic segmentation map having size as same as the input H&E image. The colored segmentation map corresponds to the legend shown at the bottom, showing the different semantic segmentation labels. The trained model can be used to generate complete segmentation map of ROIs of larger sizes by computing segmentation maps over overlapping 768×768 pixels tiles, and assigning class probabilities to the overlapping pixels by averaging.

care" class are ignored. It is essential to note that these training and testing tiles are derived from two distinct sets of images, with the testing set kept pristine and exclusively reserved for model evaluation.

Given that the images are already color-normalized, no additional color normalization is applied to these images. To standardize the images for experimentation, their pixel values are divided by 255, ensuring that values fall within the range of (0, 1).

5. METHODS

For the purpose of breast cancer tissue semantic segmentation, I propose a neural model called SPADESegResNet, drawing inspiration from SPADE[5] and ResNet[6] architectures. The goal is to leverage the powerful deep feature extraction capabilities from the residual blocks employed by ResNet, a leading model in image classification.

Originally designed for image-to-image generation, SPatially-Adaptive (DE)normalization (SPADE)[5] currently stands as a state-of-the-art model for conditional image generation. In image-to-image translation tasks, the deep neural networks based methods normally consume the input image which is then processed through series of stacks of convolution, normalization, and non-linearity layers. The normalization layers in the model tend to wash way details from the input image till the information reaches the final layer.[5] To address this issue, the SPADE model propose the spatially-adaptive normalization, a conditional normalisation layer that modulates the activations using input image through a spatially adaptive, learned transformation and can effectively propagate the input image information throughout the network.

In adapting the same concept to our semantic segmentation task, the proposed framework incorporates the power of conditional normalisation based SPADE layers, in order to effectively propagate the input H&E image information for the semantic segmentation task. The focus is on providing the input image information in the

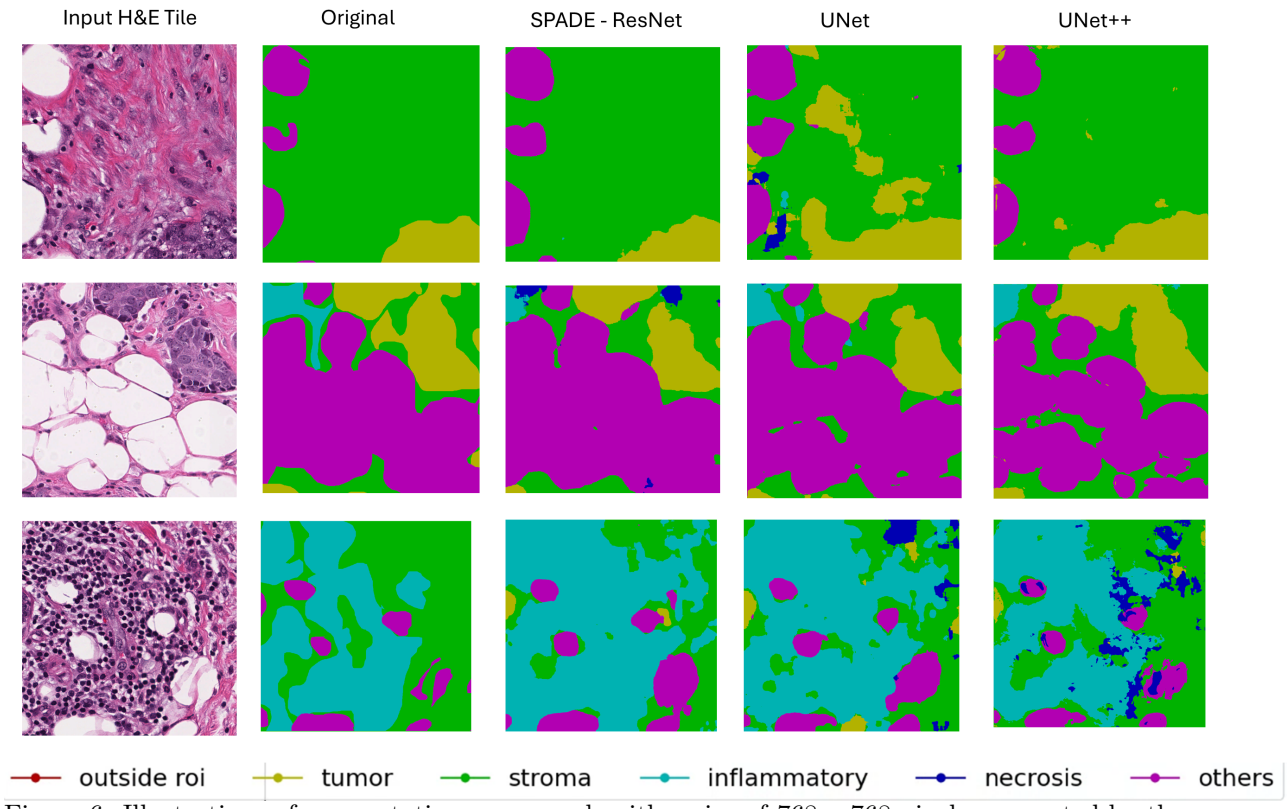

Figure 6: Illustrations of segmentation maps, each with a size of 768 × 768 pixels, computed by the proposed SPADE-ResNet alongside those produced by baseline models, including UNet and UNet++, as applied to the BCSS testing set. The colors used to indicate tissue regions are shown in the legend at bottom.

context of modulating normalised activations. The proposed framework SPADESegResNet takes input H&E image tile of size 768 × 768 pixels and outputs the same sized segmentation mask.

The proposed SPADESegResNet architecture can be visualized in Figure 5. Initially, the 768 × 768 pixel-sized tile undergoes convolutional processing before being introduced to a sequence of SPADEResNet layers. The structure of the SPADEResNet block is similar to that of the ResNet blocks employed in.[6] In the SPADE layer, the input image undergoes convolution operations to generate two modulation parameters, which are element-wise multiplied and added to the normalized activation respectively. Following a pattern similar to Batch Normalization,[9] the activation undergoes normalization in a channel-wise manner and is subsequently modulated with learned scale and bias. This better preserves the input H&E image features information against common normalization layers, safeguarding against the loss of finer details within the image. The incorporation of a deep network with residual blocks facilitates the capture of intricate features within the tissue image microenvironment. Additionally, the use of residual connections allows for the bypassing of irrelevant information, enhancing the network's ability to capture relevant details.[6]

Similar to the UNet architecture,[1] the loss function is defined by a cross-entropy loss applied between predicted pixel probabilities across all classes and actual pixel classes. The explicit formulation of the loss function is represented as follows, where N denotes the batch size, n is the total number of classes or types of tissue regions, y denotes the predicted semantic segmentation map, and $\hat{y}$ represents the actual semantic segmentation map.

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n} y_{ij} \log(\hat{y}_{ij}) \tag{1}$$

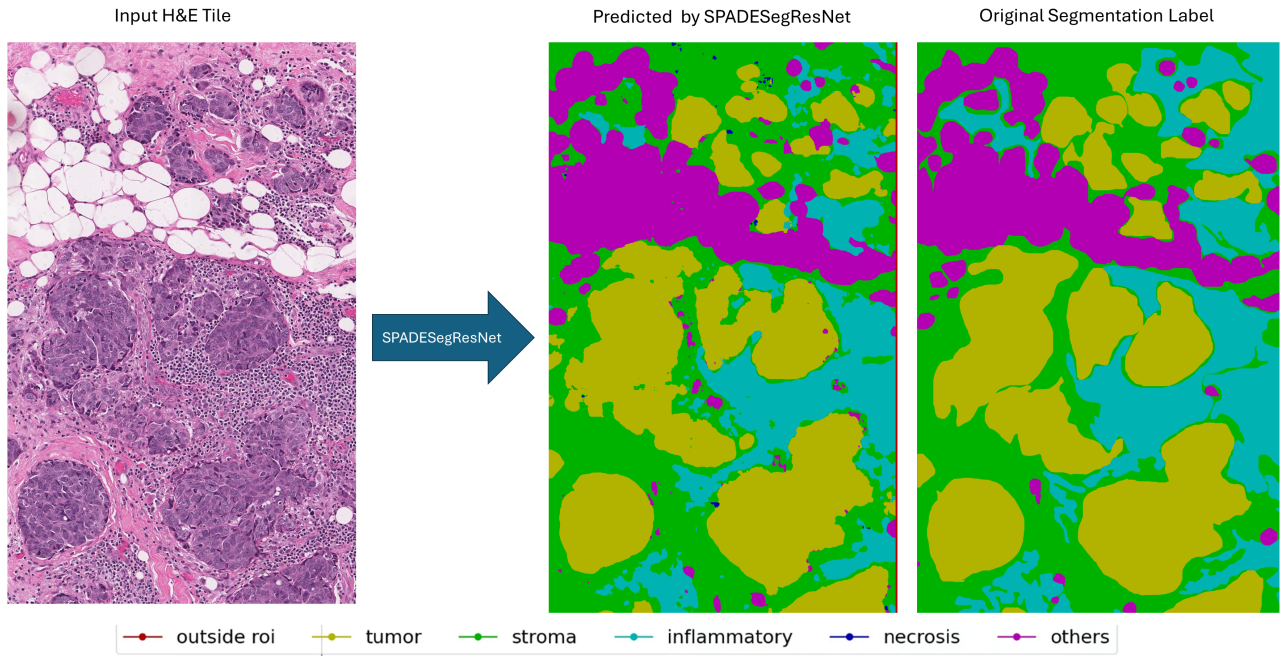

Figure 7: Complete ROI segmentation map generation of size 2756×4196 pixels by computing segmentation maps over overlapping 768×768 pixels tiles, and assigning class probabilities to the overlapping pixels by averaging. In this way, the model can be used to compute segmentation maps of larger images than those used for training.

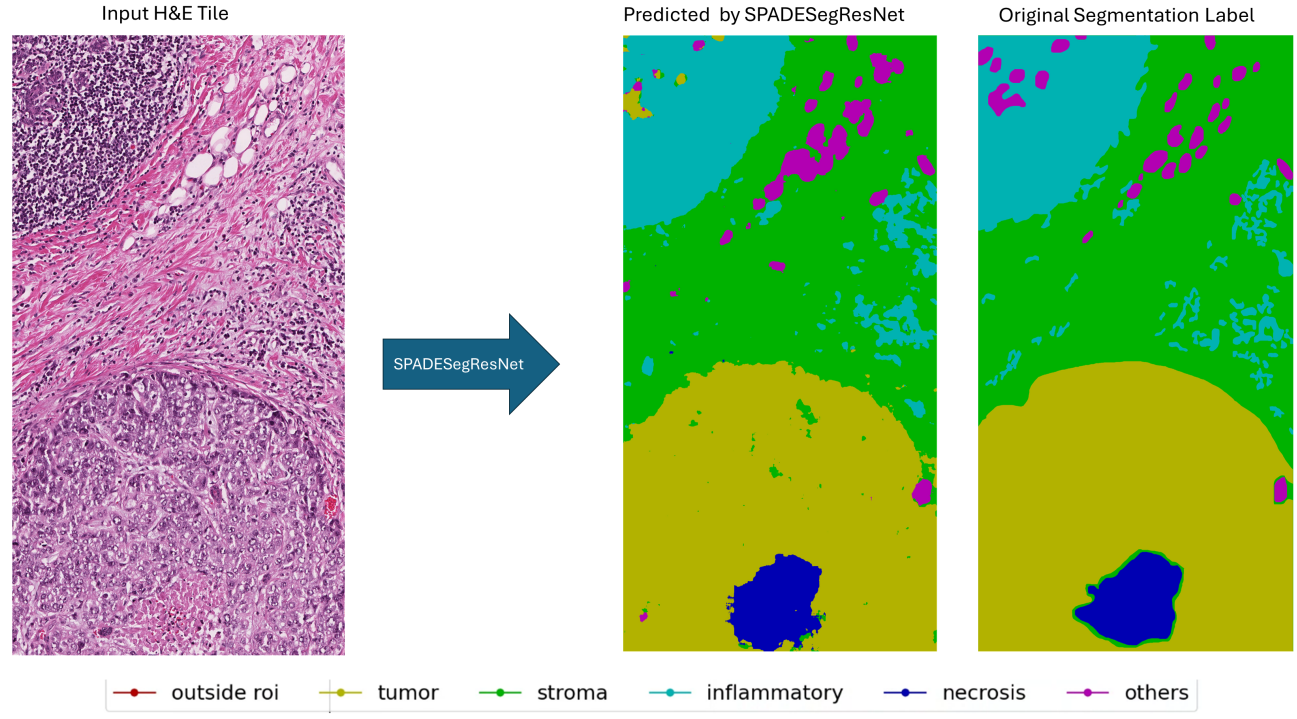

Figure 8: Complete ROI segmentation map generation of size 2233×4955 pixels by computing segmentation maps over overlapping 768×768 pixels tiles, and assigning class probabilities to the overlapping pixels by averaging. In this way, the model can be used to compute segmentation maps of larger images than those used for training.

Metric	Tissue Type	SPADESegResNet	UNet	UNet++
Dice score	Tumor	**0.77 (0.23)**	0.73 (0.22) *	0.76 (0.22)
	Stroma	0.60 (0.29)	0.60 (0.26)	**0.63 (0.28) ***
	Inflammatory	0.52 (0.29)	**0.55 (0.26)**	0.49 (0.29) *
	Necrosis	**0.50 (0.36)**	0.40 (0.31) *	0.39 (0.33) *
	Others	**0.41 (0.30)**	0.40 (0.30) *	0.37 (0.32) *
AUC-ROC (mean)	Tumor	**0.79 (0.14)**	0.78 (0.12)	0.78 (0.14)
	Stroma	0.70 (0.13)	0.71 (0.11)	**0.73 (0.13) ***
	Inflammatory	0.72 (0.17)	**0.75 (0.13) ***	0.68 (0.19) *
	Necrosis	**0.66 (0.25)**	0.64 (0.16) *	0.62 (0.21) *
	Others	0.69 (0.21)	**0.70 (0.18)**	0.63 (0.25) *
Accuracy	Overall	**0.77**	0.73	0.76

Table 1: Performance comparison of the proposed SPADESegResNet model and baseline models UNet and UNet++ in terms of metrics Dice score, mean AUC-ROC and the overall accuracy on the testing images. '*' denotes the the results obtained from the existing frameworks exhibit a statistically significant difference when compared to the results obtained from the proposed SPADESegResNet framework (without GCN).

6. EXPERIMENTS AND RESULTS

In this section, we go through the training details of the proposed framework, baseline models used for comparison, and visual and quantitative evaluation using metrics such as the Dice index score,[10] AUC-ROC, and accuracy. These metrics serve as benchmarks for evaluating the efficacy of classifying image pixels into distinct tissue types.

6.1 Experimental setup and Model Training

The whole SPADE-ResNet framework is trained on training tiles under the following specifications: a learning rate of 10^{-4}, 50 epochs, a batch size of 3, and utilization of the Adam optimizer.[11] 20% training set is held out as validation set to assess the model performance and find the optimal model. To prevent the model from overfitting, early stopping mechanism is implemented with a patience of 20. The learning rate scheduler is also incorporated which reduced the learning rate when no improvement is seen after 10 epochs.

Since the dataset is imbalanced towards some classes like necrosis, inflammatory and others, higher weights are assigned to these classes during training, ensuring a more focused model learning experience. While conventional data augmentation techniques like rotation and horizontal/vertical flipping could address the class imbalance, their implementation would substantially increase training time. Due to time constraints and the already employed weighted training approach, I thought to avoid the data augmentation in the experimentation setup.

We use UNet[1] and UNet++ or Nested UNet[3] as baseline models for comparison.

All models including the proposed and baseline models are implemented using Pytorch[†], on NVIDIA GeForce GTX TITAN X single GPU system with a dedicated GPU RAM of 12GB. Given the extensive nature of the experimental setup, involving numerous training runs, an additional Kaggle GPU P100 accelerator with 16GB GPU memory was utilized for training these models. The submission includes the complete codebase, uploaded on GitHub with restricted public access.

We train both the models on the training images using the same set of parameters as used to train the proposed SPADEResNet model. The performance of all models is evaluated on the testing images.

6.2 Results

The segmentation maps derived from testing images are illustrated in Figure 6. A careful examination highlights the exact segmentation achieved for tumor regions (yellow) by our model in contrast to the UNet and UNet++ architectures, where instances of mis-identification are observed. Additionally, our model effectively captures stroma (green), other tissue types (purple), and demonstrates fewer instances of misidentifying necrosis (dark blue) highlighting the efficacy of the SPADESegResNet framework. However, the segmentation accuracy for

[†]https://pytorch.org/

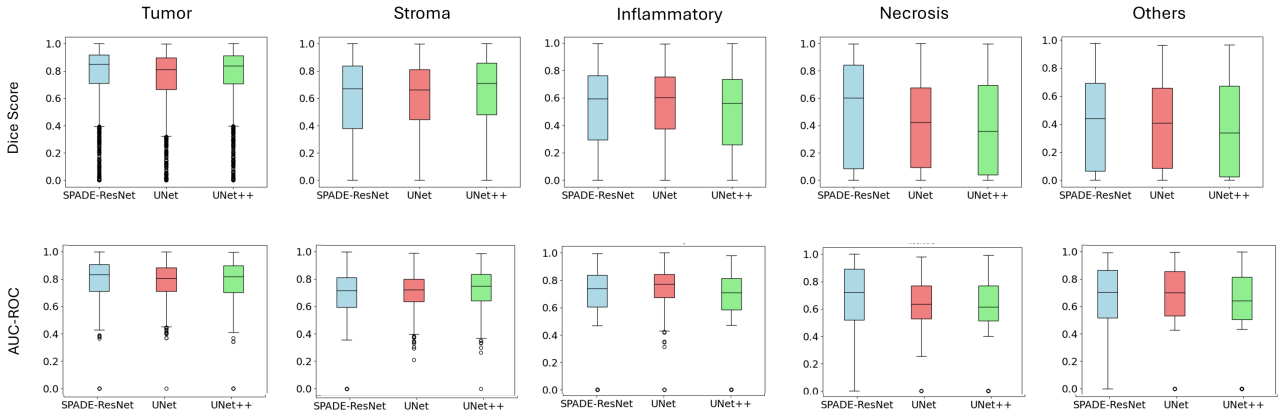

Figure 9: Boxplots showing the distribution of Dice scores and AUC-ROCs over all testing images

the inflammatory class (depicted in turquoise) appears somewhat compromised, as it is being detected instead of stroma, possibly due to the limited representation of this class in the dataset and overfitting resulting from weighted training.

The trained model can be used to generate complete segmentation map of ROIs of larger sizes by computing segmentation maps over overlapping 768×768 pixels tiles, and assigning class probabilities to the overlapping pixels by averaging. The sample generation of ROI of size 2967×4133 pixels can be visualized in Figure **??**, Figure 7 and Figure 8.

We employ three performance metrics to assess the accuracy of classifying image pixels into distinct tissue types: Dice index score[‡], AUC-ROC, and classification accuracy. The Dice score for a given tissue type is calculated as twice the size of the intersection divided by the sum of the sizes of the two sets of pixels corresponding to that particular tissue type. Both the Dice index score and AUC-ROC are computed across all testing images, and their mean and standard deviation are calculated, providing insights into the robustness of the proposed framework. Furthermore, we applied the Mann-Whitney U statistical significance test (Mann-Whitney U test).[12] This involved calculating p-values to compare the outcomes of existing frameworks with those of the proposed SPADESegResNet framework. We consider results statistically significant when the corrected p-values are below 0.05.

To determine the net accuracy, we count all correctly classified pixels across the entire testing image set and divide this count by the total number of pixels. Table 1 presents the computed metrics over the testing sets for the proposed framework and the baseline models. Additionally, to gain an understanding of the distribution of Dice scores and AUC-ROC across the images, we plot boxplots in Figure 9.

The results detailed in Table 1 and the boxplots depicted in Figure 9 clearly highlight the superior performance of SPADE-ResNet compared to the other baseline methods, particularly in the context of tissue regions such as Tumor and Necrosis, as indicated by higher Dice score and AUC-ROC. While the overall performance for Necrosis and Others surpasses that of the baseline models, it is not on par when compared with the performance observed for Tumor, primarily due to the underrepresentation of these tissue regions. UNet++ exhibits notable accuracy in identifying Stroma, but with comparatively lower metrics in other aspects.

SPADESegResNet demonstrates success in capturing tumors, achieving a Dice score of 0.77 and an AUC-ROC of 0.79. In terms of the overall accuracy score, the SPADESegResNet achieves the highest accuracy of 77%. Overall, the proposed framework demonstrates a moderately superior performance compared to the baseline models.

7. DISCUSSION

The proposed SPADESegResNet, shows robust segmentation capabilities across various tissue types, including Tumor, Stroma, and Necrosis. However, it faces challenges in achieving satisfactory performance for tissue types

[‡]https://oecd.ai/en/catalogue/metrics/dice-score

such as Inflammatory. In an effort to enhance segmentation outcomes, I tried a context-aware segmentation network which generates segmentation maps for the central portion of tissue images by utilizing the relatively larger tissue portion as input. Specifically, I tried 512×512 pixels tissue tile as input to predict a semantic mask of 256×256 pixels. Despite these efforts, the performance did not meet expectations.

An alternative approach that could be considered involves leveraging global context-aware networks like vision transformers for semantic segmentation tasks. While this technique was not explored in the current study, the decision was influenced by the observation that, during data analysis, the global context of images may not be essential for identifying specific portions of tissue images. Consequently, deep convolutional networks with limited receptive fields were seemed sufficient for the segmentation task at hand. Nevertheless, it remains a point of interest to investigate the applicability and potential benefits of vision transformers for this specific segmentation task.

An ablation study is worth conducting to assess the significance of the SPADESegResNet model's design. Unfortunately, I couldn't undertake it due to time constraints. However, judging from the performance, we cannot deny the possibility that SPADE layers are effectively conveying input image information through the neural network.

8. CONCLUSIONS

In this work, I propose the SPADESegResNet model for robust semantic segmentation of breast cancer tissue images. The model demonstrated significant performance both visually and quantitatively, as evidenced by the Dice score, AUC-ROC, and accuracy metrics. In comparison to baseline models like UNet and UNet++, the SPADESegResNet achieved moderately improved results. Additionally, we explored how the model can be utilized to compute segmentation maps for regions larger than the images used for training.

This study opens avenues for the development of robust and interpretable models for semantic segmentation tasks. Looking ahead, it points to potential areas for improvement, acknowledging the evolving landscape of deep learning architectures. As we navigate this complex landscape, the study contributes to the ongoing dialogue and lays the foundation for future advancements in computational pathology.

REFERENCES

[1] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).

[2] Long, J., Shelhamer, E., and Darrell, T., "Fully convolutional networks for semantic segmentation," 3431–3440 (06 2015).

[3] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J., [*UNet++: A Nested U-Net Architecture for Medical Image Segmentation: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*], vol. 11045, 3–11 (09 2018).

[4] Jahanifar, M., Tajeddin, N., Koohbanani, N., and Rajpoot, N., "Robust interactive semantic segmentation of pathology images with minimal user input," in [*2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*], 674–683, IEEE Computer Society, Los Alamitos, CA, USA (oct 2021).

[5] Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y., "Semantic image synthesis with spatially-adaptive normalization," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2019).

[6] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 770–778 (2016).

[7] Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A. T., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S. E., Ismail, A. F., Saad, A. M., Ahmed, J., Elsebaie, M. A. T., Rahman, M., Ruhban, I. A., Elgazar, N. M., Alagha, Y., Osman, M. H., Alhusseiny, A. M., Khalaf, M. M., Younes, A.-A. F., Abdulkarim, A., Younes, D. M., Gadallah, A. M., Elkashash, A. M., Fala, S. Y., Zaki, B. M., Beezley, J., Chittajallu, D. R., Manthey, D., Gutman, D. A., and Cooper, L. A. D., "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics* **35**, 3461–3467 (Feb. 2019).

[8] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M., "The cancer genome atlas pan-cancer analysis project," *Nature Genetics* **45**, 1113–1120 (Sept. 2013).

[9] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*Proceedings of the 32nd International Conference on Machine Learning*], Bach, F. and Blei, D., eds., *Proceedings of Machine Learning Research* **37**, 448–456, PMLR, Lille, France (07–09 Jul 2015).

[10] Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., and Kikinis, R., "Statistical validation of image segmentation quality based on a spatial overlap index1," *Academic Radiology* **11**, 178–189 (Feb. 2004).

[11] Kingma, D. and Ba, J., "Adam: A method for stochastic optimization," *International Conference on Learning Representations* (12 2014).

[12] Mann, H. B. and Whitney, D. R., "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics* **18**(1), 50 – 60 (1947).