Multi-Label Semantic Segmentation on the BCSS Dataset

Tristan Lecourtois
BioAI Research Engineer Candidate - Technical Assessment

January 2026

Abstract

This report outlines the development of a deep learning pipeline for the semantic segmentation of Breast Cancer Semantic Segmentation (BCSS) images. By implementing a U-Net architecture optimized with a hybrid Dice-Cross Entropy loss function, we demonstrate robust tissue classification across multiple biological classes. Experimental results demonstrate that the standard U-Net achieves competitive performance, often surpassing more complex architectures such as TransUNet and nnU-Net on patch-level segmentation. Patch-wise evaluations are complemented by full WSI reconstructions, highlighting U-Net's ability to maintain coherent tissue segmentation across large-scale images. We also investigate the use of a pre-trained ResNet-101 encoder to improve feature representation, although gains were limited.

1 Introduction

Breast cancer is one of the most prevalent cancers worldwide, and accurate tissue segmentation is so important for effective diagnosis and treatment planning. Traditional methods for diagnosing breast cancer depend intensely on imaging procedures such as mammography, ultrasound, and MRI, as well as tissue biopsy for definitive determination. However, they can be time-consuming, expensive, and may now and then lead to false-positive or false-negative. Semantic segmentation techniques, particularly those based on deep learning have shown promise in automating this process. Among these techniques, the U-Net architecture has gained popularity due to its encoder-decoder structure with skip connections, which allows for precise localization and efficient training even with limited data. This report explores the application of U-Net models for breast cancer tissue segmentation using the BCSS dataset. The methodology, experimental setup and results aim to provide a comprehensive understanding of the challenges and potential solutions in this domain.

2 Dataset Analysis

The Breast Cancer Semantic Segmentation (BCSS) dataset is composed of histopathology image patches extracted from whole-slide images (WSI) of breast cancer tissue originating from the Cancer Genome Atlas (TCGA). It includes over 30,000 segmentation patches of breast cancer tissue regions. Each patch has been resized to a fixed dimension of 224×224 pixels to provide a good trade-off between computational efficiency and the preservation of morphological features necessary for tissue identification at 0.25 MPP (Microns Per Pixel). Each pixel in the mask belongs to one of three semantic tissue classes:

- **Outside ROI (Class 0):** This class encompasses background regions, slide artifacts, or areas not included in the primary region of interest. In our training pipeline, these pixels are treated as "ignore labels" to prevent bias from non-tissue information.
- **Tumor (Class 1):** Represents the malignant epithelial cells. It is characterized by high cellular density and nuclear atypia.
- **Stroma (Class 2):** Includes the connective tissue and microenvironment surrounding the tumor cells.

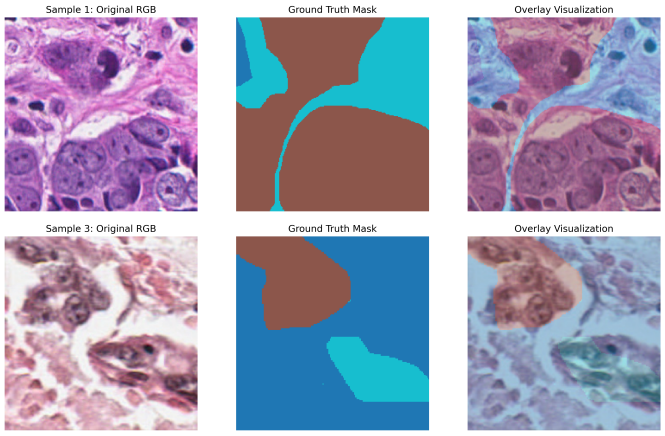

Figure 1: Two examples of histopathology image patches with their corresponding segmentation masks and overlay visualizations.

The pixel-wise class distribution in Figure 3 shows a relatively balanced dataset, with 38.16% of tumor pixels, 30.39% of stroma, and 31.45% of pixels outside the

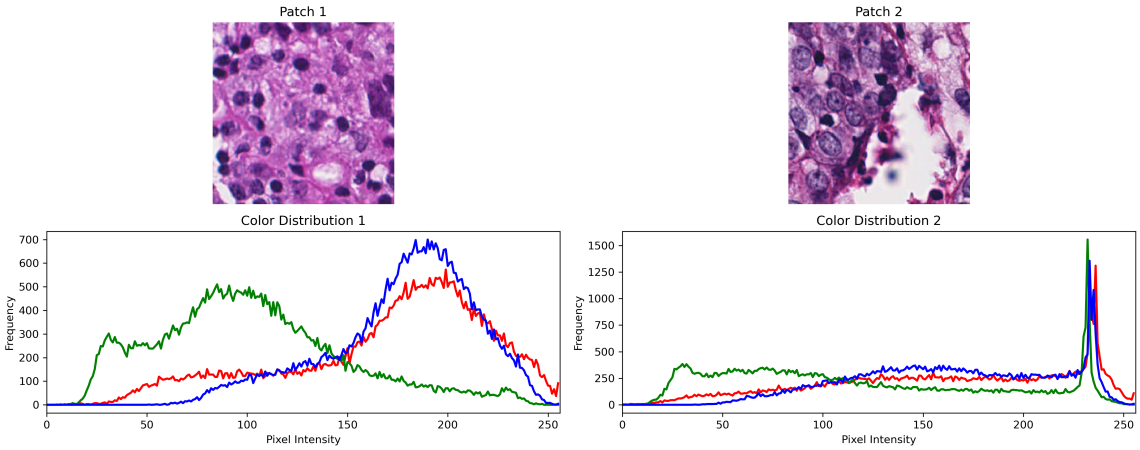

Figure 2: RGB histogram analysis of two distinct training patches. The non-overlapping peaks quantify the staining variability between samples

region of interest (ROI). Unlike typical medical segmentation datasets where pathological regions are highly underrepresented, this dataset does not suffer from severe class imbalance. This reduces the risk of model bias toward background predictions and allows the use of standard loss functions such as Cross-Entropy or Dice Loss without strong class weighting.

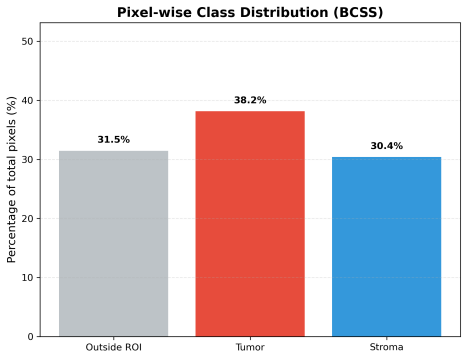

Figure 3: Pixel-wise class distribution of the dataset

Another important challenge in digital pathology is the significant variability in Hematoxylin and Eosin (H&E) staining. This "domain shift" is caused by differences in tissue preparation, reagent concentrations, and scanner hardware across clinical centers.

As demonstrated by the color histogram analysis (see Figure 2), the pixel intensity distributions for identical tissue components do not perfectly overlap between different patches. This lack of alignment suggests that a model trained without color-invariant strategies might "overfit" to specific shades of pink or purple

To address this, we integrated Color Jittering and Gaussian Blur. By artificially shifting the color histograms during training, we force the model to disregard unstable pixel intensities and focus instead on invariant morphological features.

3 Methodology

3.1 Data Pipeline & Pre-processing

To improve the model's generalization and robustness against the variability of histopathological slides, a data augmentation pipeline was implemented using the *Albumentations* library. Data augmentation is a technique used to artificially increase the size and diversity of the training dataset by applying various transformations to the original images. We applied randomly geometric transformations (rotations and flips) to account for the rotational invariance of histological tissues, alongside elastic transforms to simulate physical slide distortions. *ColorJitter* and *GaussianBlur* were used to prioritize morphological structures over color intensity. Finally, all patches were normalized using ImageNet statistics.

3.2 Models

3.2.1 U-Net

We utilized a U-Net architecture [2], which is the state-of-the-art for biomedical segmentation due to its skip-connections that allow the recovery of fine-grained cellular details during the up-sampling path. The encoder part of the U-Net comprises a series of convolutional blocks, each followed by a max-pooling layer to down-sample the feature maps. The decoder upsamples the feature maps using transposed convolutions and concenates them with the corresponding feature maps from the encoder via the skip connections. Further convolutional layers refine the upsampled feature maps to produce the final segmentation map. The bottleneck connects the encoder and decoder, applying convolutions to the lowest resolution feature maps before upsampling. A final 1x1 convolution reduces the number of output channels to the number of classes. A softmax function is applied to obtain the probability distribution for each class.

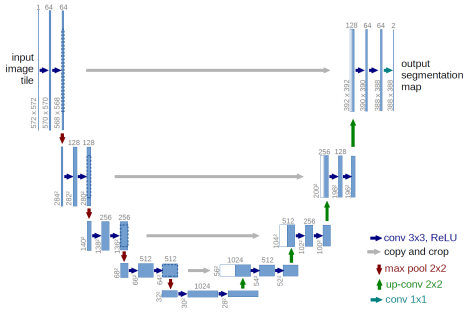

Figure 4: U-Net Architecture

3.2.2 TransUNet

To overcome the limitations of standard CNNs in capturing long-range dependencies, we implemented the TransUNet architecture [3]. While U-Net excels at localizing high-resolution details via skip connections, its purely convolutional nature limits its ability to model global context. TransUNet addresses this by adopting a hybrid design: a Transformer-based encoder handles global feature extraction while a CNN decoder ensures precise localization.

In this framework, the input image is first processed by a CNN backbone to extract high-level feature maps. these maps are then tokenized and fed into a Vision Transformer (ViT) layers. The self-attention mechanism in the Transformer allows the model to relate distant tissue structures, which is vital for identifying complex breast cancer morphologies. Specifically, the self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V represent the Query, Key, and Value matrices derived from the image patches, and d_k is the scaling factor. The attended features are then upsampled and fused with high-resolution CNN features via skip connections to produce the final segmentation mask. This synergy enables the model to leverage both the global semantic understanding of Transformers and the fine-grained spatial precision of U-Net.

3.2.3 Dense U-Net

The Dense U-Net [5] architecture incorporates the principles of DenseNet [4] into the U-Net framework. Instead of standard convolutional layers, each block consists of densely connected layers where each layer receives the feature maps of all preceding layers as input. In the context of the BCSS dataset, this architecture offers two major advantages: the concatenation of feature maps allows the model to preserve low-level histological details (like cell membrane edges) throughout the entire network. Moreover, it mitigates the vanishing gradient problem, which is crucial when training deep models on large pathology patches.

3.2.4 nnU-Net (no-new-U-Net)

The nnU-Net framework [6] represents the current state-of-the-art in medical image segmentation. Unlike other architectures that focus on novel architectural blocks, nnU-Net is a self-configuring framework that automatically adapts the entire segmentation pipeline (pre-processing, architecture, training, and post-processing) based on the dataset's properties. By utilizing a 2D U-Net baseline and focusing on automated hyperparameter tuning (such as patch size, batch size, and normalization constants), nnU-Net consistently outperforms custom-designed models on histological benchmarks.

3.3 Loss Function Selection

The choice of the loss function is essential for semantic segmentation in histopathology, as the model must achieve both pixel-wise classification accuracy and structural coherence of tissue regions. To this end, we implemented a hybrid loss function combining Categorical Cross-Entropy (CE) and Multi-class Soft-Dice Loss:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} \tag{2}$$

The Cross-Entropy loss operates at the pixel level, penalizing the model based on the negative log-likelihood of the correct class. It provides smooth gradients and ensures that each pixel is correctly classified:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3}$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted probability for class i. To handle regions outside the Region of Interest (ROI), we utilize an *ignore_ index* for Class 0, preventing the background from biasing the malignant tissue feature extraction.

While Cross-Entropy focuses on local intensity, it often fails to account for global structural overlap, especially in cases of class imbalance. We implemented a differentiable Soft-Dice Loss to directly maximize the F1-score of the predicted segments:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{p\in\Omega} \hat{y}_p y_p + \epsilon}{\sum_{p\in\Omega} \hat{y}_p + \sum_{p\in\Omega} y_p + \epsilon} \tag{4}$$

As shown in our implementation, the logits are first converted via a *Softmax* function. The masks are then one-hot encoded to compute the intersection and union across the spatial dimensions.

4 Experimental Setup

Experiments were conducted using PyTorch on NVIDIA T4 GPU using Kaggle. We employed the AdamW optimizer. The training was initialized with a learning rate of 5×10^{-3} and a weight decay of 10^{-4}.To handle potential

plateaus in the loss landscape, we integrated a ReduceLROnPlateau scheduler. We utilized a large batch size of 128. The model was trained for 15 epochs, which provided a sufficient horizon for the hybrid loss to converge without significant overfitting. Additionally, we use early stopping based on the validation loss to prevent overfitting and reduce unnecessary training epochs.

5 Result Analysis

5.1 Quantitative Metrics

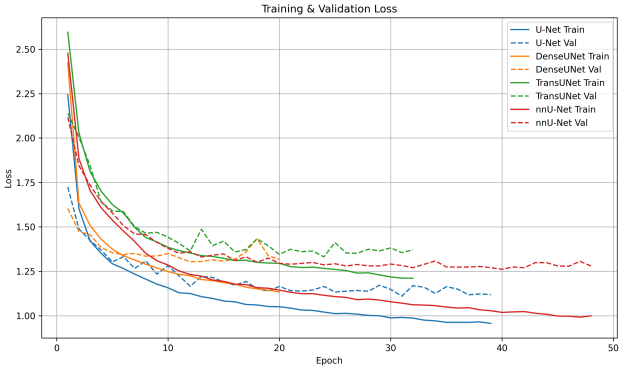

Figure 5: Training and validation loss curves for all four models.

To quantitatively evaluate the segmentation performance, we report both the Mean Intersection over Union (mIoU) and the Dice Similarity Coefficient (DSC). Figure 5 presents the training and validation loss curves for the four evaluated models: U-Net, TransUNet, Dense U-Net, and nnU-Net. All models exhibit stable convergence behavior. Interestingly, the standard U-Net converges faster and reaches a lower training loss than both TransUNet and nnU-Net. This behavior may appear counterintuitive, as nnU-Net is often considered a state-of-the-art framework for medical segmentation. However, this difference can be explained by the higher model complexity of Transformer-based and self-configuring architectures, which typically require longer training schedules and larger datasets.

Figure 6 shows the evolution of the Mean IoU on both training and validation sets. The IoU metric evaluates the spatial overlap between predicted and ground truth regions:

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{5}$$

All architectures achieve consistent improvements in IoU over training. TransUNet and U-Net achieve the highest final IoU scores which can reflect their aibility to capture both global tissue context and fine-grained boundaries.

Figure 6: Mean IoU on training and validation sets for all models.

While IoU provides a strict overlap measure, the Dice Similarity Coefficient (DSC) is preferred in medical segmentation because of its sensitivity to small anatomical structures:

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \tag{6}$$

It measures the harmonic mean between precision and recall at the pixel level. It is suitable for histopathology segmentation where precise boundary delimitations is very important. Unlike accuracy, Dice is robust to class imbalance and emphasizes correct region overlap rather than background dominance.

Model	Dice Score (DSC)
U-Net	0.834
Dense U-Net	0.812
TransUNet	0.823
nnU-Net	0.809

Table 1: Dice Similarity Coefficient for the four evaluated models.

5.2 Qualitative Assessment

To evaluate whether using a more powerful encoder improves segmentation results, we replaced the standard encoder of our U-Net with a ResNet-101 pre-trained on ImageNet. The goal was to leverage the representations already learned on a large dataset in order to capture richer features and potentially improve segmentation accuracy on histological patches. However, as shown in Figure 7, the results did not improve as much as expected, and the segmentation quality remained comparable to the standard U-Net.

In addition to patch-wise evaluation, we also wanted to assess the segmentation performance directly on whole-slide images (WSIs) by comparing the standard U-Net with TransUNet. As shown in Figure 8, U-Net performs

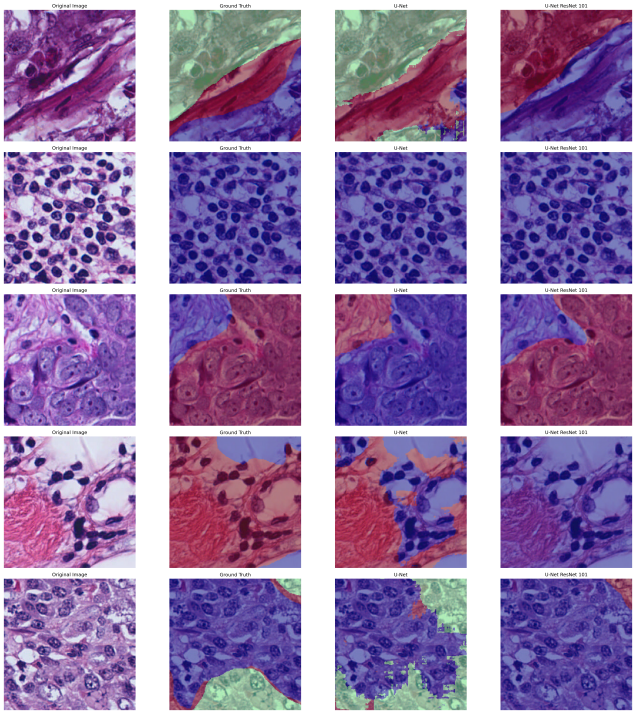

Figure 7: Comparison of segmentation results with U-Net with a ResNet-101 encoder.

better than TransUNet on this WSI. To achieve this, we first split the WSI into smaller patches, performed segmentation on each patch individually, and then reconstructed the full WSI by stitching the patch predictions back together.

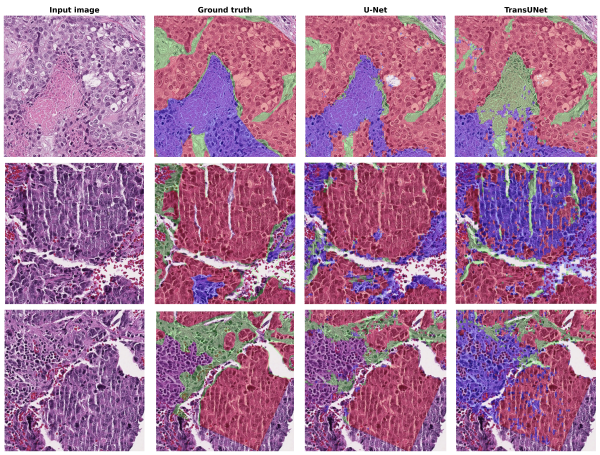

Figure 8: Whole-slide image (WSI) segmentation comparison

6 Discussion

Experimental results demonstrate that a standard U-Net architecture can achieve strong performance on the BCSS dataset. However, models were trained for a relatively small number of epochs due to computational and time constraints. More complex architectures such as TransUNet and nnU-Net typically require longer training schedules to fully converge, especially when Transformers or self-configuring pipelines are involved. Extending the training duration would likely allow these models to better optimize their parameters.

In addition, the implemented U-Net variants were intentionally kept simple, with a limited number of convolutional layers and moderate channel sizes. This design choice was motivated by GPU memory limitations and the need for fast experimentation. However, increasing the depth of the network or using wider convolutional filters could improve the model's capacity to capture complex histopathological patterns, such as subtle tumor–stroma boundaries and heterogeneous tissue structures.

Beyond architectural improvements, alternative segmentation paradigms could also be explored. One direction is the use of the Segment Anything Model (SAM). SAM is a foundation model trained on a large-scale dataset of images and segmentation masks using a prompt-based mechanism. Instead of directly predicting semantic labels, SAM generates segmentation masks based on user prompts such as points, bounding boxes, or coarse masks. In the context of histopathology, SAM could be adapted by providing tumor or stroma prompts on a few representative regions of a slide. These prompts could guide the model to produce high-quality segmentation masks without requiring full pixel-level annotations. This approach would significantly reduce annotation costs while still enabling large-scale tissue segmentation.

Another important direction involves Self-Supervised Learning (SSL), particularly with methods such as DINO (Self-Distillation with No Labels). DINO trains vision transformers by enforcing consistency between different augmented views of the same image, allowing the model to learn meaningful visual representations without any annotations. By pretraining a backbone network on large collections of unlabeled histopathology slides, the model could learn domain-specific features such as cellular morphology, texture patterns, and tissue organization. These learned representations could then be fine-tuned for segmentation with only a small number of labeled samples, reducing the dependence on expensive expert annotations.

7 Conclusion

Our work provides a solid baseline for breast cancer segmentation. Future work will explore more advanced training strategies and foundation models to further improve accuracy while reducing annotation requirements.

References

[1] Amgad M, et al. (2019). *Structured crowdsourcing enables convolutional segmentation of histology images.*

Bioinformatics.

[2] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, arXiv preprint arXiv:1505.04597, 2015.

[3] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*, arXiv preprint arXiv:2102.04306, 2021.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely Connected Convolutional Networks*, arXiv preprint arXiv:1608.06993, 2018.

[5] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, *Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network*, *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 6, pp. 1275–1285, Jun. 2020, doi:10.21037/qims-19-1090.

[6] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*, arXiv preprint arXiv:1809.10486, 2018.

Appendix: Additional Figures

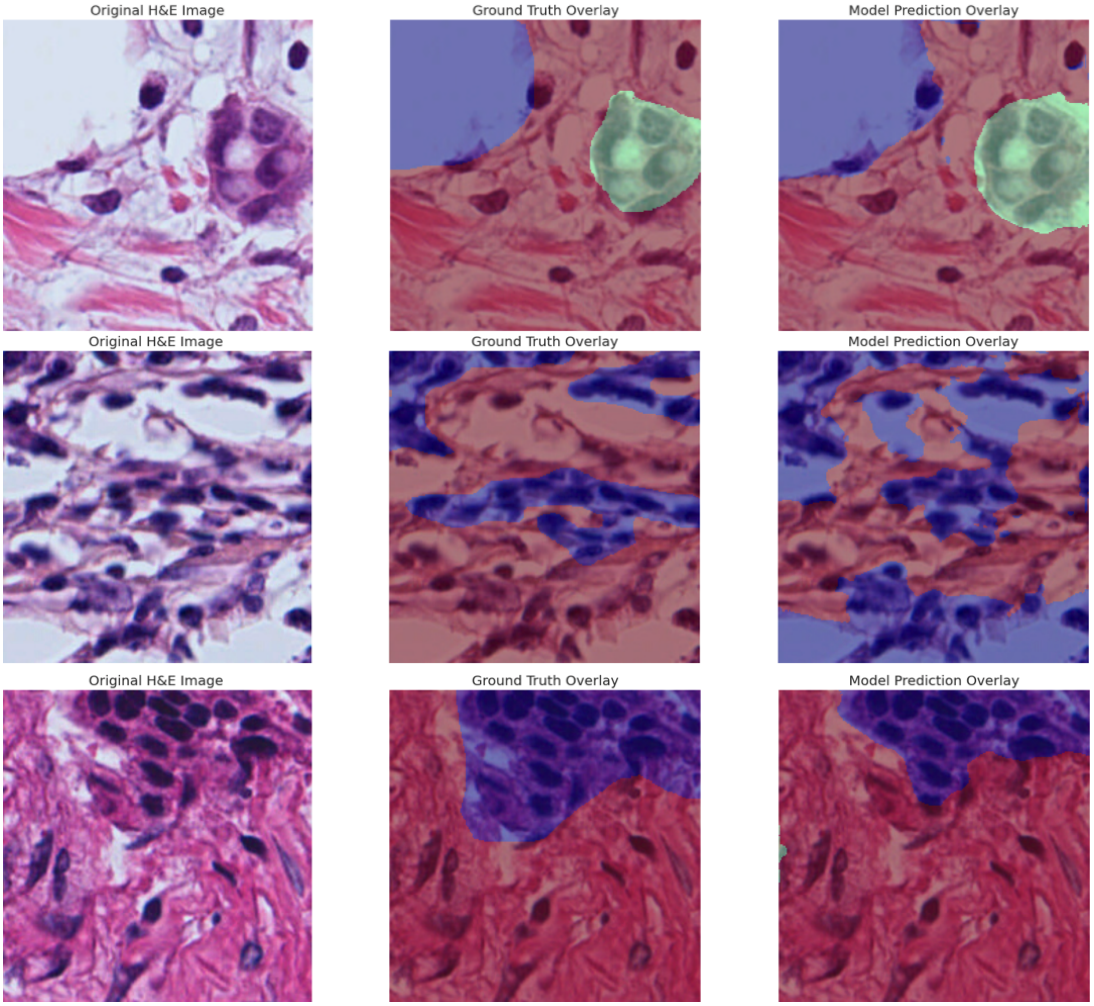

Figure 9: From left to right: original histopathology image, ground truth mask overlay, predicted segmentation by U-Net.