Joint Breast Neoplasm Detection and Subtyping using Multi-Resolution Network Trained on Large-Scale H&E Whole Slide Images with Weak Labels

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Breast cancer is the most commonly diagnosed cancer and the use of artificial intelligence (AI) to help diagnose the disease from digital pathology images has the potential to greatly improve patient outcomes. However, methods for training these models for detecting, segmenting, and subtyping breast neoplasms and other proliferative lesions often rely on costly and time-consuming manual annotation, which can be infeasible for large-scale datasets. In this work, we propose an weakly supervised learning framework to jointly detect, segment, and subtype breast neoplasms. Our approach leverages top-k multiple instance learning to train an initial neoplasm detection backbone network from weakly-labeled whole slide images, which is then used to automatically generate pixel-level pseudo-labels for whole slides with only one subtype. A second network is trained using these pseudo-labels, and slide-level classification is performed by training an aggregator network that fuses the embeddings from both backbone networks. We trained and validated our framework on large-scale datasets with more than 100k whole slide images and demonstrate its effectiveness on tasks including breast neoplasms detection, segmentation, and subtyping.

Keywords: Digital Pathology, Multiple Instance Learning, Breast Neoplasm

1. Introduction

Breast cancer has surpassed lung cancer as the most commonly diagnosed cancer worldwide, with 2.26 million new cases in 2020. Early detection and accurate diagnosis is crucial for determining an optimal treatment strategy and improving patient outcomes. Medical institutions have only recently begun the transition from microscopes to digital pathology workflows using high-resolution whole-slide images (WSIs), which introduces tremendous opportunity to incorporate modern AI methods to more thoroughly and systematically help detect cancer, provide segmentation masks to aid in interpretation, and subtyping the detected cancer regions to aid in treatment decision-making. Training AI algorithms for digital pathology is challenging due to the large size of WSIs and limited availability of ground truth data. Recent systems break WSIs into small tiles and aggregate tile information for slide-level predictions (Campanella et al., 2019). However, ground truth data is often only available at the case or specimen level, making it labor-intensive and expensive to create pixel or object level ground truth at the scale needed for industry training. Most studies in the field have only focused on small-scale datasets such as (Litjens et al., 2018; Zuley et al., 2016), which may not reflect the complexity and variability of real-world distributions, making it difficult to determine the impact of algorithmic differences when large-scale datasets are available.

© 2023 Anonymous.

Here, we propose an AI system that detects and subtypes breast neoplasms at both pixel and slide level granularity without relying on region annotations. AI-assisted cancer detection has proven value in clinical settings (Raciti et al., 2022), and automated subtyping has the potential to standardize clinical protocol and improve the accuracy of detecting invasive cancers. However, it requires high accuracy to avoid biasing pathologists and eroding trust in the AI system. To address these challenges, a framework is needed to leverage weak ground truth extracted from clinical diagnostic reports automatically, scaling training and evaluation to large-scale datasets that represent the distribution of cases seen in clinical use with sufficiently large sample sizes. Such a framework could greatly improve the efficiency and accuracy of AI-based diagnosis in digital pathology and ultimately improve patient outcomes.

Contribution: In this paper, we propose a joint breast cancer detection, segmentation, and subtyping framework that does not rely on manual annotations for training. Our approach decomposes the system training into 3 stages. In stage 1, it uses top-k multiple instance learning (Campanella et al., 2019) to train an initial neoplasm detection backbone network, which is then used to generate pseudo-labels for regions of interest. In stage 2, a second network is trained to subtype the image regions at a coarse granularity using pseudo-labels. In stage 3, slide-level classifications are obtained by a per-slide aggregator network that fuses the embeddings from the stage 1 and 2 networks. Our method is trained and validated using 100K+ H&E WSIs. Segmentation performance is evaluated using 105 exhaustively annotated WSIs with different cancer subtypes. Additionally, we compute the RoI-level classification accuracy on the BRACS dataset (Brancati et al., 2021) without any finetuning. We further evaluate the specimen level neoplasm detection performance on 7k+ slides and investigate different options of obtaining the WSI-level prediction results.

2. Related work

Despite the lack of large-scale datasets, there have been many pioneering studies that have explored the use of deep learning systems on breast hematoxylin and eosin (H&E) WSIs (Chan et al., 2023). For example, in Cruz-Roa et al. (2017), a convolutional neural network (CNN) was employed for detecting invasive tumors on WSIs. Similarly, Gandomkar et al. (2018); Mi et al. (2021) used a deep network to classify WSIs as benign or cancerous, and then further categorized cancer cases into subtypes. Le et al. (2020) employed deep CNNs to quantitatively assess tumor and tumor-infiltrating lymphocytes, and Wang et al. (2022) explored the use of deep learning models to directly assist in the histological grading of breast cancer. These studies demonstrate the potential of utilizing deep learning techniques to improve the diagnostic accuracy and efficiency of breast cancer detection and analysis. Pati et al. (2022) explored using a hybrid graph network to classify the breast neoplasm into fine-grained subtypes and evaluated on the BACH (Aresta et al., 2019) and BRACS datasets (Brancati et al., 2021). A few studies have attempted to employ pseudo-labeling techniques, similar to those used in our study for mapping breast digital pathology diagnosis reports to pixel-level ground truth, in order to grade prostate cancer (Bulten et al., 2020; Marini et al., 2021; Silva-Rodriguez et al., 2021).

Recent studies have begun to investigate the use of multi-resolution inputs for convolutional neural networks (CNNs) in digital pathology analysis. This approach, which involves

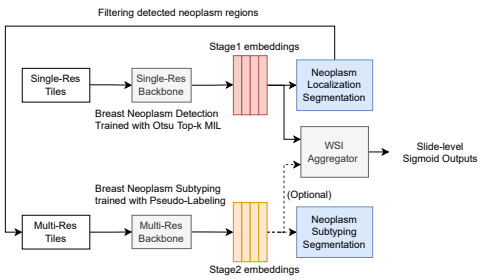

Figure 1: Framework overview

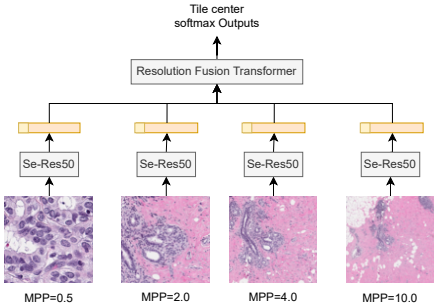

Figure 2: Illustration of the multi-res backbone architecture.

providing the CNN with different scales or resolutions of the same image, such as multiple crops or down-sampled versions, enables the CNN to capture both fine-grained details at high resolution and contextual information at lower resolutions. For instance, in some studies (Li et al., 2021; van Rijthoven et al., 2021; Sandbank et al., 2022), multi-resolution inputs of whole slide images (WSIs) were utilized to improve the classification performance of self-supervised CNNs by exploring the information at different scales. Similarly, in Karimi et al. (2020), the authors employed an ensemble of three different CNNs to enhance the accuracy of predicting prostate Gleason scores. Additionally, in Chen et al. (2022), the authors expanded the receptive field of tile networks with a hierarchical vision transformer trained with self-supervised learning.

3. Methods

Our proposed framework consists of three successive networks shown in Figure 1. The first CNN is trained using a top-K MIL framework (Campanella et al., 2019) on specimen level labels to detect neoplasm tiles in slides. This CNN is then used for selecting subtyping training tiles from subsets of the training slides that each uniquely contain one specific breast neoplasm. The slide-level ground truth labels are used as pseudo-labels for the selected subtyping training tiles. We then train a multi-resolution CNN on these tiles using the pseudo-labels. Both CNNs are used as tile-level feature extractors, and an attention-based network aggregates and fuses these embeddings to produce final slide-level predictions. The ground truth used in this study was automatically parsed from the clinical diagnosis reports, containing 6 common breast neoplasms at specimen level: Invasive breast carcinoma of no special type (previously known as invasive ductal carcinoma or IDC), Invasive lobular carcinoma (ILC), Ductal carcinoma in situ (DCIS), Lobular carcinoma in situ (LCIS), Atypical ductal hyperplasia (ADH), and Atypical lobular hyperplasia (ALH) (Lakhani et al., 2012). The specimen is labeled as benign if none of the 6 subtypes and other rare cancer types were reported.

3.1. Breast Neoplasm Detection Trained with Otsu Top-k MIL

We first train a SE-ResNet-50 (Hu et al., 2019) using top-k multiple instance learning (MIL) to detect non-benign tiles in a slide. Top-k MIL is commonly used to train models on WSIs

when only the slide-level diagnosis is available (Campanella et al., 2019). The foreground regions of the WSI are extracted using an HSV filter with fixed HSV ranges tuned for non-fat breast tissues in H&E WSIs. At the start of each training epoch, the CNN first scores each foreground tile extracted at 0.5 μm / pixel (MPP), 20X in general, with the cancer subtype presence probabilities with sigmoid outputs. For each training step, the CNN weights are only updated using the top k highest ranked instances for each neoplasm subtype. Though the tile-wise ground truth labels for the non-benign slides can be noisy, the top k tiles picked by CNN on benign slides are reliable. The optimization is thus mainly driven by the benign slides and was shown to be effective when training with large-scale datasets (Campanella et al., 2019). To adapt to lesions of varying sizes, we used the Otsu's method (Otsu, 1979) on the tile predictions to dynamically choose the k for each slide while setting a maximum cap $k_{max} = 400$ fixed in our experiments. After training, a rough segmentation map can be generated for each subtype by thresholding the network sigmoid outputs.

3.2. Breast Neoplasm Subtyping trained with Pseudo-Labeling

While the detection CNN can detect non-benign tiles, it does not always provide mutually exclusive scores for subtypes of breast neoplasms. To address this issue, we train a second CNN that is specialized in subtyping the neoplasm tiles detected by the detection CNN. We selected a subset of the training dataset where only one lesion type presents in each slide, and assumed that all neoplasm tiles share the same subytpe. In practical clinical reports, many atypia labels could be missing when more severe findings present. Thus, we merged the in situ and atypia subtypes and formulated the breast subtyping problem as a 4-class classification problem involving IDC, ILC, DCIS/ADH, or LCIS/ALH. The second CNN is trained to predict the mutually exclusive probabilities that a tile center is within a subtype. We used the detection network to generate pseudo-segmentation masks to filter all neoplasm tiles and invasive cancer tiles on the selected slides. The subtyping network is then trained to classify the tile center of the filtered tiles into 4 classes.

It is important to consider larger contexts in breast neoplasm subtyping, since pathologists often zoom in and out to check the lesions boundaries and other macro-scale morphological patterns for breast cancer diagnosis. To address this issue, we use of a simple yet effective multi-resolution CNN network, as illustrated in Figure 2. We extract multiple tiles of different sizes at the same tile center, which are then rescaled to the same size. Each tile is processed separately by a SE-ResNet-50 without weight sharing. The resulting embeddings are treated as tokens and fused by a small transformer network with fixed positional encoding (Vaswani et al., 2017). The network output is a softmax function, assuming that the tile center class labels are mutually exclusive. 2D coarse-grained subtype labelmaps can be derived from the network outputs using $argmax$ of the tile-level subtype probabilities.

3.3. Attention based aggregator for slide-level prediction

Given that the detection and subtyping backbone networks can each provide feature embeddings for WSI tiles, we use an aggregator network to fuse the embeddings and obtain the final slide-level predictions. Since the presence of each neoplasm is not mutually exclusive at slide-level by nature, the network outputs a binary prediction for each class. We explored two approaches: (1) training an aggregator network using only the tile embedding from the

Table 1: The summary of the in-house dataset used in this study showing number of slides and specimens. We also break down the number of specimens for neoplasm subtypes and benign according to the clinical diagnosis.

Split	Slides	Specimens	IDC	ILC	DCIS	LCIS	ADH	ALH	Benign
Train	93369	44683	5209	905	9724	2968	3245	2812	28262
Validation	24020	7235	813	154	1542	432	508	429	4655
Test	7989	2000	272	51	453	151	147	151	1222
Total	125378	53918	6294	1110	11719	3551	3900	3392	34139

detection network; (2) or training an aggregator network to fuse the tile embeddings from both the detection and subtyping networks. For the latter, we only extract subtyping embeddings on tiles detected with breast neoplasm to avoid tiles that are out-of-distribution for the subtyping network. When there is no neoplasm detected, a zero vector is used as a placeholder. We use all tissue tiles for extracting detection embeddings. Section 5.3 details the network architecture choices. The slide-level outputs are activated with a sigmoid function. To map the slide-level predictions to the specimen level ground truth, we take the maxpool of the sigmoid outputs across all the slides in the same specimen.

4. Data

The data used in this study contains over 125k WSIs with mixed biopsies and excision specimens, and were acquired from a large cancer center as described in Table 1. The majority of the slides were scanned using Leica scanners. The patients were anonymized and randomly sampled from real clinical cohorts to represent a natural patient population. The train and validation splits were created at specimen level with random sampling therefore sharing the similar distribution. The specimen ground truth labels were obtained by automatically parsing the diagnostic reports. The validation set was used for hyper-parameter tuning and model selection. We reserved 2000 specimens for the test set used for performance reporting. We randomly sampled 188 slides from the non-benign test slides that were exhaustively annotated by board certified pathologists with labelled polygons to analyze segmentation performance. We reserved 83 as a tuning set for operating point selection and 105 for evaluation. As an external dataset, we also acquired the entire BRACS dataset (Brancati et al., 2021) to evaluate the RoI based breast subtyping performance without finetuning on BRACS labels.

5. Results

5.1. Breast Neoplasm Localization

To evaluate the localization of all breast neoplasm detection and invasive breast cancer, we obtained the all neoplasm and invasive segmentation masks from the backbone networks as described in Section 3.1 and Section 3.2. The invasive breast cancer segmentation can be obtained from either the detection or subtyping network. In Figure 3, we show the box plots of per-slide dice, recall and precision of 105 test WSIs. In summary, the invasive cancer segmentation (S2 Inv.) from the sub-typing network has better average dice (0.66 ± 0.24) than the invasive cancer segmentation (S1 Inv.) from the detection backbone (0.56 ± 0.30).

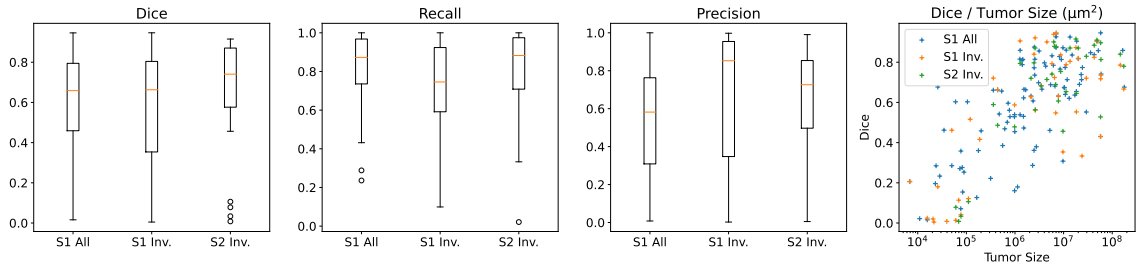

Figure 3: The plots of pixel-wise dice, recall and precision of the cancer segmentation including all breast neoplasms (S1 All) or only the invasive cancer areas (S1 Inv. and S2 Inv.). All the dice scores are plotted with the tumor size on the right.

Table 2: Ablative analysis for the components of the breast neoplasm subtyping network.

Architecture	MPP=0.5	MPP=2.0	MPP=4.0	MPP=10.0	PE	Tile-Wise Accuracy
SE-ResNet-50×1	✓	✗	✗	✗	✓	0.74
SE-ResNet-50×2	✓	✓	✗	✗	✓	0.78
SE-ResNet-50×3	✓	✓	✓	✗	✓	0.77
SE-ResNet-50×4	✓	✓	✓	✓	✗	0.79
SE-ResNet-50×4	✓	✓	✓	✓	✓	**0.81**

It also surpassed the dice scores (0.56 ± 0.30) of the all neoplasm segmentations (S1 All) from the detection network, although the areas of invasive cancer are larger. As a reference, we also show that larger tumors tend to have higher dice scores in general on the right of Figure 3.

5.2. Breast Neoplasm Subtyping

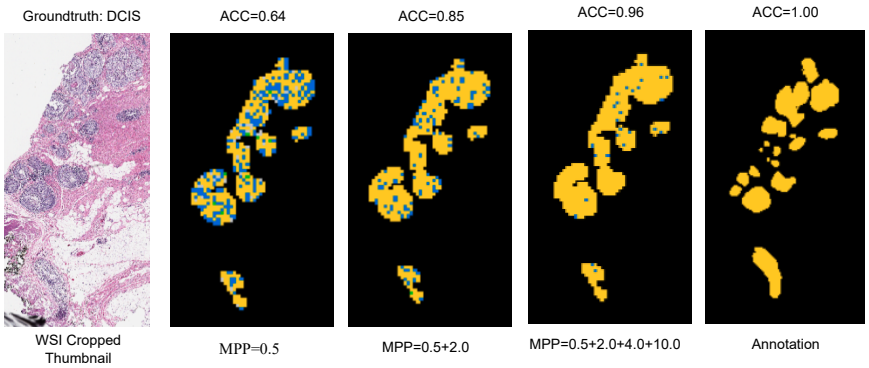

Figure 4: Visual examples showing a WSI crop where adding more contexts progressively improved the tile-wise subtyping tile-wise accuracy (ACC). Different colors in the segmentation labelmaps indicate different breast neoplasm subtype categories.

An ablative analysis was performed on 105 annotated slides with mixed breast neoplasm subtypes to investigate the contributions of different components of the proposed neoplasm sub-typing network as shown in Table 2. Since the neoplasm region areas vary within and

Table 3: The F1 scores computed on the the BRACS dataset's RoIs.

Split	Network	Training method	Normal	Benign	Atypia	In-Situ	Invasive
Test	CNN 0.5 mpp (Pati et al., 2022)	Fully Supervised	0.420	0.423	22.67	50.33	0.770
Test	MS-CNN 1.0+0.5+0.25 mpp (Pati et al., 2022)	Fully Supervised	0.503	0.443	31.67	57.33	0.860
Test	HACT-Net (Pati et al., 2022)	Fully Supervised	0.615	0.474	40.42	66.44	0.884
All	MS-CNN 10.0+4.0+2.0+0.5 mpp	Weakly supervised	0.876		0.685		0.731
Test	MS-CNN 10.0+4.0+2.0+0.5 mpp	Weakly supervised	0.902		0.528		0.826

between WSIs, we computed the tile-wise classification accuracy (ACC) within each slide and averaged it among the whole dataset as the performance indicator. By adding larger contexts, we observed an increase of ACC from 0.74 to 0.81 except adding $MPP = 4.0$ on top of higher resolutions did not show improvement. By removing the position encoding, which provides the ordering of resolutions to the resolution fusion transformer, we noticed the accuracy dropped from 0.81 to 0.79. As shown in Table 3, we also benchmarked the subtyping network on both the test splits and the entire BRACS dataset (Brancati et al., 2021) without finetuning. The required context size $10 \times 10\mu m^2$ is mostly larger than the annotated BRACS RoIs. To predict the RoI classification labels, we located each RoI in the corresponding WSI and extracted the context window needed by our network at the center of the previous RoI. If the RoI is not within the cancer detection segmentation, it is then classified as Normal / Benign. We quoted the F1 scores from Table 5 of Pati et al. (2022) as reference. Though we had a slightly different label grouping setup compared to Pati et al. (2022), the F1 scores of Invasive cancers can be directly compared. On the test set, though our proposed network was trained with weakly supervised pseudo-labelling, it reached an F1 score of 0.826 similar to fully supervised CNNs (0.770 to 0.860). Though not directly comparable, we provide fine-grained subtyping numbers from Pati et al. (2022) for reference. The proposed framework achieved a high F1 score (0.902) for the Normal / Benign because our framework decomposes the tile classification problem into detection and subtyping separately.

5.3. Breast Neoplasm Detection

The specimen level AUC scores for detecting each breast neoplasm subtype are shown in Table 4. As a baseline, we take the maximum tile scores from the detection network to represent the specimen level prediction (Maxpool MIL). With the tile embedding from the detection network only, we investigated the attention-based MIL aggregators similar to Attention MIL (Ilse et al., 2018) and DSMIL (Li et al., 2021). We also evaluated a few aggregator architectures as illustrated in Figure 5 to fuse the tile embedding from both the detection and subtyping networks. In each architecture, the outputs of the attention branches are concatenated and then fed through the output layer. For the DSMIL based architectures we experimented with the source of the critical instances which are used as the query vectors in the attention layers. The critical instances are determined by selecting the tiles with the highest probability for each class. We tried three approaches: (1) using the detection embeddings to determine critical instances where the selected embeddings would then be used as the query vectors for attention on both the detection and subtyping embeddings; (2) using the subtyping embeddings in the same manner; (3) using both detection and subtyping embeddings to select two sets of critical instances to be used independently in their own attention layers. We notice an improvement in AUC compared to the Maxpool

MIL baseline for all aggregators except in benign AUC while the largest improvement is seen in ADH AUC (0.896 vs 0.951). The performance on invasive carcinomas is similar between the two overall approaches (≥0.983 AUC for all learned aggregators), however we found that inclusion of the subtyping embeddings can improve AUC on the non-invasive lesions, specifically ADH (0.951), DCIS (0.982), and LCIS (0.965).

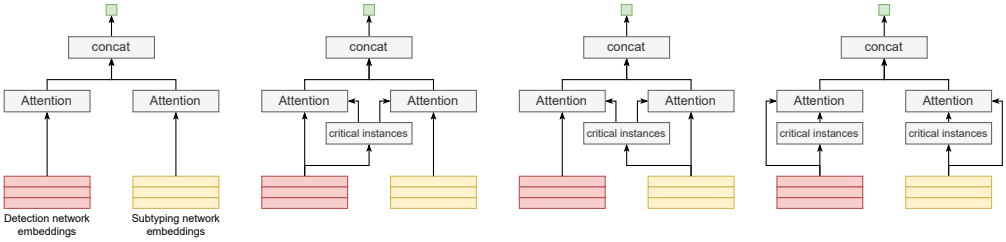

Figure 5: Architectures for fusing detection and subtyping embeddings. From left to right: 1) attention MIL, 2) DSMIL with detection embeddings as critical instances, 3) DSMIL with subtyping embeddings as critical instances, and 4) DSMIL with separate critical instances.

Table 4: The specimen level detection AUCs for different subtypes breast neoplasm.

Name	ADH	ALH	DCIS	LCIS	IDC	ILC	Inv.	Benign	Inv. vs. Other
Positive	147	151	453	151	272	51	387	1206	387
Total	1617	1553	1950	1695	1931	1931	2000	2000	794
Detection Embedding									
Maxpool MIL	0.896	0.946	0.951	0.932	0.976	0.971	0.976	**0.986**	0.945
Attention MIL	0.950	**0.969**	0.980	0.962	**0.986**	**0.994**	**0.984**	0.984	0.962
DSMIL	0.943	0.965	0.978	0.963	0.985	0.989	0.983	0.982	0.958
Detection Embedding + Subtyping Embedding									
Attention MIL	**0.951**	0.966	**0.982**	0.961	0.984	0.992	0.983	0.984	0.955
DSMIL (cross attention 1)	0.938	0.962	0.978	**0.965**	**0.986**	**0.994**	**0.984**	0.979	0.961
DSMIL (cross attention 2)	0.941	0.959	0.973	0.952	0.985	0.993	**0.984**	0.981	**0.963**
DSMIL (separate attention)	0.950	0.965	**0.982**	0.959	0.984	**0.994**	0.983	0.983	0.956

6. Conclusion

In this study, we proposed a framework for jointly detecting, segmenting, and subtyping breast neoplasms, which requires no manual annotation for training. We trained and validated the framework using a large-scale dataset of over 100k WSIs. Results showed that the proposed subtyping network improved the segmentation of invasive cancer regions and that using a multi-resolution network improved tile-level cancer classification accuracy. Our weakly supervised subtyping network achieved similar F1 scores compared to fully supervised CNNs. We evaluated slide-level detection AUCs using different aggregator architectures, observing an improvement over using the detection network alone, though the improvement was marginal when using both detection and subtyping embedding.

References

Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, Jack Zeineh, Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, Maximilian Baust, Quoc Dang Vu, Minh Nguyen Nhat To, Eal Kim, Jin Tae Kwak, Sameh Galal, Veronica Sanchez-Freire, Nadia Brancati, Maria Frucci, Daniel Riccio, Yaqi Wang, Lingling Sun, Kaiqiang Ma, Jiannan Fang, Ismael Kone, Lahsen Boulmane, Aurélio Campilho, Catarina Eloy, António Polónia, and Paulo Aguiar. BACH: Grand challenge on breast cancer histology images. 56:122–139, 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.05.010. URL https://www.sciencedirect.com/science/article/pii/S1361841518307941.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for BReAst carcinoma subtyping in h&e histology images, 2021. URL http://arxiv.org/abs/2111.04740.

Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated gleason grading of prostate biopsies using deep learning. 21(2):233–241, 2020. ISSN 14702045. doi: 10.1016/S1470-2045(19)30739-9. URL http://arxiv.org/abs/1907.07980.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. 25(8):1301–1309, 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1.

Ronald CK Chan, Chun Kit Curtis To, Ka Chuen Tom Cheng, Tada Yoshikazu, Lai Ling Amy Yan, and Gary M Tse. Artificial intelligence in breast cancer histopathology. 82(1):198–210, 2023. ISSN 0309-0167, 1365-2559. doi: 10.1111/his.14820. URL https://onlinelibrary.wiley.com/doi/10.1111/his.14820.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. pages 16144–16155, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Scaling_Vision_Transformers_to_Gigapixel_Images_via_Hierarchical_Self-Supervised_Learning_CVPR_2022_paper.html.

Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie N. C. Shih, John Tomaszewski, Fabio A. González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. 7(1):46450, 2017. ISSN 2045-2322. doi: 10.1038/srep46450. URL https://www.nature.com/articles/srep46450.

Ziba Gandomkar, Patrick C. Brennan, and Claudia Mello-Thoms. MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. 88:14–24, 2018. ISSN 0933-3657. doi: 10.1016/j.artmed.2018.04.005. URL https://www.sciencedirect.com/science/article/pii/S0933365717305031.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. URL http://arxiv.org/abs/1709.01507.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018. URL http://arxiv.org/abs/1802.04712.

Davood Karimi, Guy Nir, Ladan Fazli, Peter C. Black, Larry Goldenberg, and Septimiu E. Salcudean. Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. 24(5):1413–1426, 2020. ISSN 2168-2208. doi: 10.1109/JBHI.2019.2944643.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

SR Lakhani, IO Ellis, SJ Schnitt, PH Tan, and MJ van de Vijver. *WHO Classification of Tumours of the Breast*. 2012. ISBN 9789283224334. URL https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-The-Breast-2012.

Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Luke Torre-Healy, Richard A. Moffitt, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, Tianhao Zhao, Arvind Rao, Alison L. Van Dyke, Ashish Sharma, Erich Bremer, Jonas S. Almeida, and Joel Saltz. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. 190(7):1491–1504, 2020. ISSN 0002-9440. doi: 10.1016/j.ajpath.2020.03.012. URL https://www.sciencedirect.com/science/article/pii/S0002944020301887.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. pages 14318–14328, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Dual-Stream_Multiple_Instance_Learning_Network_for_Whole_Slide_Image_Classification_CVPR_2021_paper.html.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. 7(6):giy065, 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy065. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6007545/.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs, math].

Niccolò Marini, Sebastian Otálora, Henning Müller, and Manfredo Atzori. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. 73: 102165, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102165. URL https://www.sciencedirect.com/science/article/pii/S1361841521002115.

Weiming Mi, Junjie Li, Yucheng Guo, Xinyu Ren, Zhiyong Liang, Tao Zhang, and Hao Zou. Deep learning-based multi-class classification of breast digital pathology images. 13: 4605–4617, 2021. ISSN 1179-1322. doi: 10.2147/CMAR.S312608.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. 9(1):62–66, 1979. ISSN 2168-2909. doi: 10.1109/TSMC.1979.4310076.

Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, Maurizio Di Bonito, Giuseppe De Pietro, Gerardo Botti, Jean-Philippe Thiran, Maria Frucci, Orcun Goksel, and Maria Gabrani. Hierarchical graph representations in digital pathology. 75:102264, 2022. ISSN 1361-8415. doi: 10.1016/j.media.2021.102264. URL https://www.sciencedirect.com/science/article/pii/S1361841521003091.

Patricia Raciti, Jillian Sue, Juan A. Retamero, Rodrigo Ceballos, Ran Godrich, Jeremy D. Kunz, Adam Casson, Dilip Thiagarajan, Zahra Ebrahimzadeh, Julian Viret, Donghun Lee, Peter J. Schüffler, George DeMuth, Emre Gulturk, Christopher Kanan, Brandon Rothrock, Jorge Reis-Filho, David S. Klimstra, Victor Reuter, and Thomas J. Fuchs. Clinical validation of artificial intelligence-augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. 2022. ISSN 1543-2165. doi: 10.5858/arpa.2022-0066-OA.

Judith Sandbank, Guillaume Bataillon, Alona Nudelman, Ira Krasnitsky, Rachel Mikulinsky, Lilach Bien, Lucie Thibault, Anat Albrecht Shach, Geraldine Sebag, Douglas P. Clark, Daphna Laifenfeld, Stuart J. Schnitt, Chaim Linhart, Manuela Vecsler, and Anne Vincent-Salomon. Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. 8(1):1–11, 2022. ISSN 2374-4677. doi: 10.1038/s41523-022-00496-w. URL https://www.nature.com/articles/s41523-022-00496-w.

Julio Silva-Rodriguez, Adrian Colomer, Jose Dolz, and Valery Naranjo. Self-learning for weakly supervised gleason grading of local patterns. 25(8):3094–3104, 2021. ISSN 2168-2208. doi: 10.1109/JBHI.2021.3061457.

Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. 68:101890, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101890. URL https://www.sciencedirect.com/science/article/pii/S1361841520302541.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Y. Wang, B. Acs, S. Robertson, B. Liu, L. Solorzano, C. Wählby, J. Hartman, and M. Rantalainen. Improved breast cancer histological grading using deep learning. 33 (1):89–98, 2022. ISSN 0923-7534. doi: 10.1016/j.annonc.2021.09.007. URL https://www.sciencedirect.com/science/article/pii/S0923753421044860.

Margarita L. Zuley, Rose Jarosz, Bettina F. Drake, Danielle Rancilio, Aleksandra Klim, Kimberly Rieger-Christ, and John Lemmerman. The cancer genome atlas prostate adenocarcinoma collection (TCGA-PRAD), 2016. URL https://wiki.cancerimagingarchive.net/x/tgpp. Type: dataset.

Appendix A. Breakdown of the Test Dataset

Table 5: Statistics on the in-house test dataset used in this study showing number of specimens for each neoplasm subtype broken down by specimen type.

Specimen type	Benign	ADH	ALH	DCIS	LCIS	IDC	ILC	Microinv.	Inv.	Total
Biopsy	311	18	19	84	19	76	7	2	101	637
Excision	815	127	128	347	127	182	39	8	265	2038
Total	1126	145	147	431	146	258	46	10	366	2675

We provide the label statistics for our in-house test dataset in Table 5 to extend Table 1. Note that the total specimen count shown is 2675 though the total number of unique specimens is 2000 since multiple breast neoplasms can co-exist in the same specimen.

Appendix B. Network architectures

B.1. Detection network

For the SE-ResNet-50 (Hu et al., 2019) detetion network used in Section 3.1, we change the last convolutional layer to output 512 channels instead of the original 2048 with a projection layer. The final SE block was also modified to retain the constant reduction ratio of $r = 16$ throughout the network. Lastly, we replaced the final global averaging pooling layer with max pooling.

B.2. Subtyping network

We use four separate SE-ResNet-50 networks, one for each magnification level, to extract features of size 2048. These networks follow the original SE-ResNet-50 implementation without the modifications used in the detection network. We use a standard Transformer

network (Vaswani et al., 2017) with 3 encoder layers. We create the classifier input by flattening the output of the Transformer to obtain a single embedding of size 8192.

B.3. DSMIL variants

The DSMIL architecture used only the detection embedding follows the architecture described in (Li et al., 2021). The modified architecture variants that use both detection and subtyping embedding are presented below.

Let $\mathbf{x}_d \in \mathbb{R}^{T\times 512}$ be our 512-dimensional detection embedding for all T tissue tiles, and let $\mathbf{x}_s \in \mathbb{R}^{N\times 8192}$ be the 8192-dimensional subtyping embeddings for all N tiles predicted to be neoplastic by the detection network. For the first variant, we use the detection embedding for critical instance selection. The detection embedding are fed through a linear layer with 7 output units corresponding to the classes ADH, ALH, DCIS, LCIS, IDC, ILC, and all invasive, i.e.,

$$\mathbf{i} = \sigma(\mathbf{W}_i \mathbf{x}_d), \tag{1}$$

where σ is the logistic sigmoid function. The output of this branch is connected to a binary cross entropy loss for optimizing tile-level predictions. Max-pooling the tile-level probabilities leads to selecting the embedding with the highest probability for each class which are considered the critical instances $\mathbf{c}$ such that $\mathbf{c} \subseteq \mathbf{x}$. The critical instances are then used as queries in scaled dot product attention with the full set of detection embedding and subtyping embedding in separate attention heads:

$$\begin{aligned}
\mathbf{q}_d &= \tanh(\mathbf{W}_a(\mathrm{relu}(\mathbf{W}_e\mathbf{c})) & \mathbf{q}_s &= \tanh(\mathbf{W}_a(\mathrm{relu}(\mathbf{W}_c\mathbf{c})) \\
\mathbf{k}_d &= \tanh(\mathbf{W}_a\mathrm{relu}(\mathbf{W}_e\mathbf{x}_d)) & \mathbf{k}_s &= \tanh(\mathbf{W}_a\mathrm{relu}(\mathbf{W}_e\mathbf{x}_s)) \\
\mathbf{v}_d &= \mathbf{W}_v\mathbf{x}_d & \mathbf{v}_s &= \mathbf{W}_v\mathbf{x}_s \\
\mathbf{z}_d &= \mathrm{softmax}(\frac{\mathbf{q}_d\mathbf{k}_d^T}{\sqrt{d_{k_d}}})\mathbf{v}_d & \mathbf{z}_s &= \mathrm{softmax}(\frac{\mathbf{q}_s\mathbf{k}_s^T}{\sqrt{d_{k_s}}})\mathbf{v}_s
\end{aligned}$$

where despite the shared notation for the weight matrices, the weights are not shared between the attention heads. For all models we set $\mathbf{W}_a$ to have 64 output units and all other weight matrices are square except for the output layers. The slide-level class predictions are obtained by concatenating the output of the attention followed by a linear layer and a sigmoid function:

$$\mathbf{b} = \sigma(\mathbf{W}_b[\mathbf{z}_d; \mathbf{z}_s]). \tag{2}$$

For the second variant, where we use the subtyping embedding for critical instances, the roles of x_d and x_s would simply be swapped in the notation above. The third version of the network uses both detection and subtyping embedding to compute two sets of critical instances which requires calculating two set of tile-level predictions:

$$\begin{aligned}
\mathbf{i}_d &= \sigma(\mathbf{W}_{id}\mathbf{x}_d) \\
\mathbf{i}_s &= \sigma(\mathbf{W}_{is}\mathbf{x}_s)
\end{aligned} \tag{3}$$

where we max-pool with respect to each class to select the critical instances in each source of embedding $\mathbf{c}_d$ and $\mathbf{c}_s$ such that $\mathbf{c}_d \subseteq \mathbf{x}_d$ and $\mathbf{c}_s \subseteq \mathbf{x}_s$. The critical instances are the

queries for independent attention heads along with their respective source embedding. The output of the network follows Equation (2). We also apply dropout ($p = 0.25$) to the value vectors in the attention heads and on the concatenated attention head outputs as well.

Appendix C. Optimization details

C.1. Detection network

The model was trained for 20 epochs. The training dataset consisted of 938 million unique tiles of size 224×224 taken from the $20\times$ magnification level (0.5 μm/px) of the WSIs. We applied random image augmentations to the tiles including rotations, vertical/horizontal flips, color-jittering, sharpening, blurring, and contrast stretching. We used Adam optimizer (Kingma and Ba, 2017), learning rate of 4e-5, and effective batch size of 1536.

C.2. Subtyping network

We applied random image augmentations to the tiles including vertical/horizontal flips, color-jittering, sharpening, and contrast stretching. We used Adam optimizer (Kingma and Ba, 2017), learning rate of 1e-5, and effective batch size of 128.

C.3. Aggregators

The labels for our dataset were parsed from clinical diagnosis reports which summarize findings at the specimen level. Due to this, along with memory constraints caused by the large-scale nature of breast WSIs, we used a two-pass training algorithm. We first make a forward pass without gradients, over all slides in a given specimen. We then take the slide with the highest probability as the representative slide for the specimen and compute a second forward pass with gradients computed. All aggregator models were trained using AdamW optimizer (Loshchilov and Hutter, 2019), learning rate of 5e-4, and per-device batch size of 1 with gradient accumulation tuned according the number of devices to achieve a consistent effective batch size of 32. We used early stopping monitoring the validation loss.

Appendix D. Example Breast Neoplasm Segmentation Visualizations

In accompaniment to Figure 3, which presents the localization numbers for breast neoplasm and invasive cancer segmentation, we provide a visual demonstration of the framework's performance through examples. Figure 6 depict cases where the framework achieved a Dice score higher than 0.8, indicating successful capture of the entire annotated region with minimal false positives. However, as seen in Figure 7 and Figure 8, the framework can also exhibit typical error modes, such as missing small cancer regions or over-segmentation of the biopsy site changes, which often leads to false positives.

To further illustrate the capabilities of the framework, we present two examples of invasive cancer segmentation in Figure 9 and Figure 10. These examples demonstrate the difference between the segmentation obtained from the detection network (S1 Inv.) and that obtained from the subtyping network (S2 Inv.). Through this comparison, it becomes evident that the subtyping network is able to improve both the sensitivity and specificity of the invasive cancer segmentation, as observed in our results.

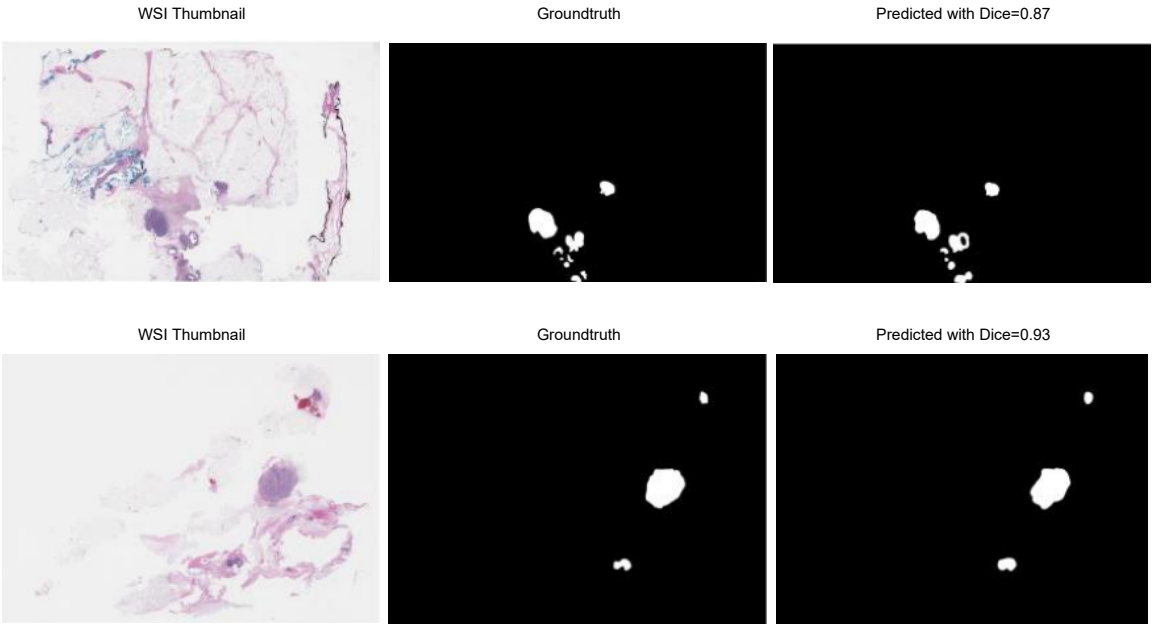

Figure 6: Example breast neoplasm segmentation with high dice scores.

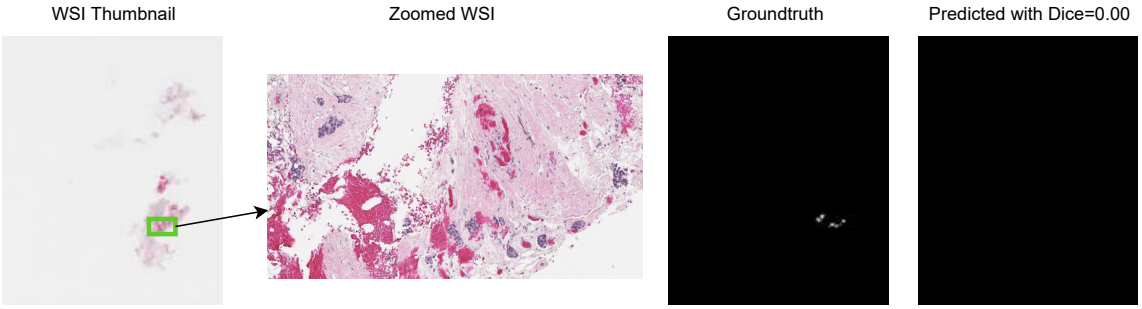

Figure 7: The only occurrence of a breast neoplasm segmentation with 0 dice score in our results, which likely occurred due to the small size of the cancer region.

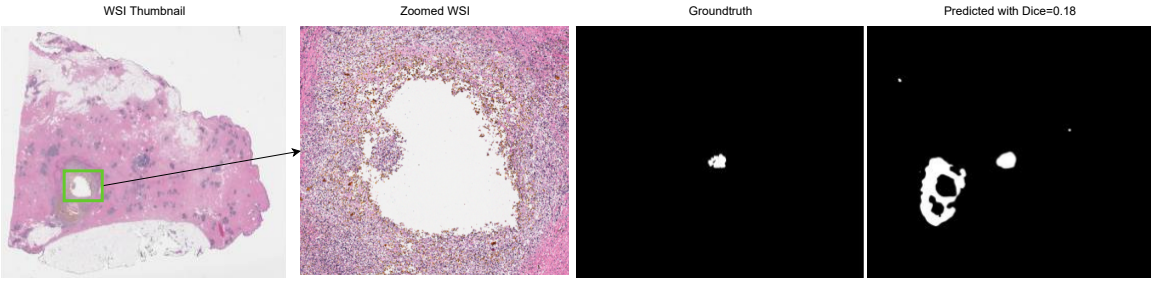

Figure 8: Example breast neoplasm segmentation with 0.18 dice score. The network over-segmented with biopsy site.

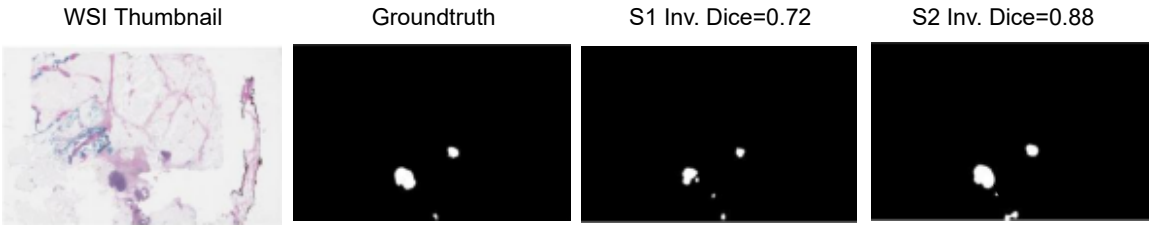

Figure 9: Example of invasive cancer segmentation where the invasive cancer segmentation generated from the subtyping network (S2 Inv.) has higher sensitivity and dice score compared to the one obtained from the detection network (S1 Inv.).

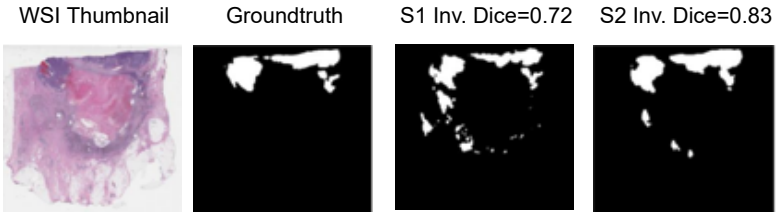

Figure 10: Example of invasive cancer segmentation where the invasive cancer segmentation generated from the subtyping network has less false positives.