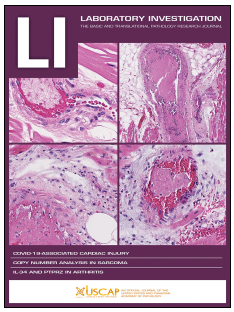



# AI-assisted detection of breast cancer lymph node metastases in the post-neoadjuvant treatment setting

Tony Xu, Dina Bassiouny, Chetan Srinidhi, Michael Sze Wai Lam, Maged Goubran, Sharon Nofech-Mozes, Anne L. Martel

PII: S0023-6837(25)00031-5

DOI: https://doi.org/10.1016/j.labinv.2025.104121

Reference: LABINV 104121

To appear in: *Laboratory Investigation*

Received Date: 25 October 2024

Revised Date: 14 February 2025

Accepted Date: 19 February 2025

Please cite this article as: Xu T, Bassiouny D, Srinidhi C, Lam MSW, Goubran M, Nofech-Mozes S, Martel AL, AI-assisted detection of breast cancer lymph node metastases in the post-neoadjuvant treatment setting, *Laboratory Investigation* (2025), doi: https://doi.org/10.1016/j.labinv.2025.104121.

CS2024401606

# AI-assisted detection of breast cancer lymph node metastases in the post-neoadjuvant treatment setting

Tony Xu[1,†], Dina Bassiouny[2,3], Chetan Srinidhi[3], Michael Sze Wai Lam[4], Maged Goubran[1,3,5,6], Sharon Nofech-Mozes[2,3,*], Anne L. Martel[1,3,*]

[1] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[2] Department of Laboratory of Medicine and Pathology, University of Toronto, Toronto ON Canada
[3] Physical Sciences Platform, Sunnybrook Research Institute, Toronto, Ontario, Canada
[4] Department of Biomedical Engineering, University of Waterloo, Waterloo ON Canada
[5] Hurvitz Brain Sciences, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada
[6] Harquail Centre for Neuromodulation, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada

[†] Corresponding author: tonylt.xu@mail.utoronto.ca
[*] Equal contribution

# Abstract

Lymph node assessment for metastasis is a common, time-consuming, and potentially error prone pathologist task. Past studies have proposed deep learning (DL) algorithms designed to automate this task. However, none have explicitly evaluated the generalizability of these algorithms to lymph nodes in breast cancer patients who have received post-neoadjuvant systemic therapy (NAT).

In this study, we create a large, 1027-slide dataset containing exclusively post-NAT breast cancer patients with detailed pathologist labels. We develop an interpretable DL pipeline to carry out two tasks: firstly, to classify slides as positive or negative for metastases, and secondly, to create a detailed, patch-level heatmap for probability of metastasis. We evaluate this pipeline with and without post-NAT treatment effect in training data, and investigate its performance relative to both slide- and patch-level tasks. We find that the presence of post-NAT treatment effect training data is relevant for both tasks, with particular benefits in pipeline specificity.

With the post-NAT testing cohort, we found that our final pipeline obtained 0.986 area under the receiver operating characteristic curve (AUC) for slide-level classification, and 70.9% specificity when calibrating for 100% sensitivity. We additionally perform an interpretability study on the outputs of our pipeline, and find that the patch-level heatmap was successful in efficiently guiding pathologists towards detecting and correcting erroneous predictions that were made with an uncalibrated network.


# Introduction

With recent advances in scanning, storage, and computational capabilities, fully digitized pipelines for analysis of whole-slide histopathologic images (WSIs) have become feasible, encouraging interest in their deployment to the clinic. Artificial intelligence (AI) plays a crucial role in automating these analytical processes, particularly in routine and labor-intensive tasks. One such task is the assessment of lymph nodes by pathologists to exclude metastatic spread in surgical lymph node specimens. This is integral to cancer staging using criteria such as the tumor, node, metastasis (TNM) staging system. However, the complexity of this procedure can render it laborious, time-consuming, and error-prone, especially when considering cases with minimal metastatic deposits or when the cancer cells have morphologic features overlapping with lymphocytes or histiocytes, particularly without stromal desmoplasia. Promoted by the availability of large, densely annotated public WSI repositories, many computer-aided WSI analysis methods have been proposed for reducing pathologist workload, expediting turnaround time, and acting as a second reader [1–3].

In patients who have received neoadjuvant systemic therapy (NAT), the response to treatment in the axillary lymph nodes – specifically based on the number of involved nodes and size of metastatic deposits – is an important determinant of survival post-NAT. This information can be used in conjunction with other quantitative metrics derived from resected breast tissue to evaluate the residual cancer burden (RCB) [4]. Although the RCB has been shown to provide useful prognostic information [5] and has been recommended internationally since 2018, routine clinical practice resists adopting the metric due to how laborious it is to evaluate [6]. Although many published papers review the use of AI models to detect LN

metastasis in breast cancer [2,7,8], most of these have either focused on the well-known CAMELYON datasets [9], where none of the patients received chemotherapy prior to resection, or have used large internal datasets where the number of post-NAT cases is not specified [10]. In the latter study, the authors obtained very strong results despite utilizing only a slide level label; however, they did not explicitly evaluate the distribution shift between chemo-naive and post-NAT patients. Notably, they found that twelve WSIs (four false negatives and eight false positives) contained tissue alterations resulting from NAT during their assessment of prediction errors. Other studies have explored generalizability towards separate cancer types [3,11], or axillary lymph nodes [12] but have not explored the impact of alterations attributed to therapy response such as lymphocytic depleted area, fibrosis, and sheets of foamy histiocytes – characteristics termed as “treatment effects”. These features are not commonly found in chemo-naive patients, and therefore may confound models trained solely on this patient group. False positive results due to such effects have been mentioned in several studies [10,12,13] but there has been no large-scale assessment of AI model accuracy for lymph node metastases detection in the neoadjuvant setting. To address this gap and investigate the generalizability of DL algorithms to post-NAT patients, we created a large internal dataset we term Post-NAT Lymph Nodes (Post-NAT-LN). This dataset is comparable in size to the CAMELYON Challenge dataset, and contains detailed pathologist annotations for metastatic deposits, normal lymph node tissue, and post-NAT treatment effects.

For breast cancer, nodal staging involves investigating the presence of macrometastasis (>2mm diameter), micrometastasis (>0.2mm and <2mm diameter) and isolated tumor cells (ITCs) (<0.2mm diameter). The clinical relevance of ITCs is debated [14–17], but Wong et al. suggests that ITCs may be more clinically relevant in the post-NAT setting [14]. Proper detection of ITCs is also key when determining RCB, where their presence may greatly affect the size of the tumor bed in the lymph node. For these reasons, our Post-NAT-LN training and testing datasets include WSIs that only contain ITCs, and these are labeled as positive cases.

To perform automated metastasis detection on the CAMELYON and Post-NAT-LN datasets, we formulate a deep learning (DL) pipeline that jointly outputs a prediction of metastasis presence on a slide-level, and an interpretable detailed heatmap of tumor presence. The former output provides a second reader to evaluate the WSI, and the latter has the potential to enhance pathologist interpretability and streamline the initial read of the slide. We take the publicly released CAMELYON challenge dataset [9] as a chemo-naive dataset, and Post-NAT-LN as a post-NAT dataset to evaluate the relevance of treatment effect regions for slide-level prediction accuracy and patch-level heatmap quality. We also explore the effect of training a model with and without post-NAT treatment effect regions on the overall *generalizability* of the pipeline.

Finally, we evaluate the benefits of an interpretable, two-stage tumor detection pipeline. A pathologist will be asked to view the top ten patches with the highest tumor probability values output by the patch-level heatmap and evaluate the accuracy of the slide-level prediction based on the contents of these ten patches. We use this study to show the importance of having accurate and detailed heatmaps for pathologist interpretability and for minimizing errors in slide-level classification.

To summarize our contributions, we:

1. Perform the first investigation on the importance of post-NAT data for generalizability of DL networks.
2. Probe the trade-offs when using post-NAT data to train networks on the tasks of creating a heatmap of suspicious regions and obtaining a slide-level classification for presence of metastasis in lymph nodes.
3. Demonstrate the importance of accurate heatmaps for minimizing prediction errors and improving interpretability by simulating a pathologist workflow.

Our proposed two-stage pipeline could significantly benefit pathologist workflows by directing prioritization to analytically complex or ambiguous slides, increasing sensitivity to small metastatic deposits, and speeding up analysis via heatmap guidance. The findings of this study could also elucidate the significance of having post-NAT patient data in training datasets when developing DL networks. Thus, we hope to inform future work investigating the creation and clinical deployment of DL algorithms for detecting metastasis in lymph nodes. Code used to train and evaluate our pipeline will be released at https://github.com/martellab-sri/lymph-node-metastases-detection.

# Methods and Materials

Due to the computational complexity associated with passing an entire WSI through an AI model, we employed the commonly-used multiple instance learning (MIL) deep learning paradigm in this work. MIL involves first breaking a WSI into smaller image patches and feeding them through a pre-trained and frozen feature extractor. This compresses the information present inside each patch into a “feature vector” which can be acted on using MIL methods to aggregate into a slide level prediction. Contrary to standard practice which uses feature extractors trained to classify natural images [18,19], or extractors that are trained using self-supervised methods [20,21], our two-stage pipeline first trains a feature extractor to predict patch-level tumor presence and uses the extracted patch features to create a detailed tumor probability heatmap. Then, patch features are combined using MIL aggregation techniques to yield a slide-level prediction for presence of metastasis. A workflow for the overall method can be found in Figure 1.

## Datasets

### CAMELYON

The Cancer Metastasis in Lymph Nodes Challenge (CAMELYON) dataset is a large, publicly available dataset containing sentinel lymph nodes resected from chemo-naive BrCa patients [9]. The slides were stained using routine Haematoxylin and Eosin (H&E) stain and several brands of slide scanners were used to digitize the images (20x magnification,0.23 - 0.25 μm/pixel). We combined datasets from both CAMELYON16 and CAMELYON17, and used the CAMELYON16 test dataset as our test set. Overall, the dataset consists of 377 patients and 898 WSIs, which was split on patient-level with 769 slides (292 positive, 477 negative) in the training set, and 129 (49 positive, 80 negative) in the testing set.

The CAMELYON16 training dataset was exhaustively annotated for metastatic deposits (all tumor areas are fully annotated), and any WSIs containing *only* ITCs were removed. The CAMELYON17 dataset provided detailed annotations for a subset of 50 WSIs in the training set, which were exhaustively annotated for micrometastases and macrometastases, but not ITCs. The CAMELYON17 dataset *included* slides that only contain ITCs. All slides were also labeled on the slide-level as either positive or negative for the presence of cancer, and there are no slides that only contain ITCs in the test set.

### Post-NAT-LN

This dataset was collected at the Sunnybrook Health Sciences Centre, Toronto, Canada, with approval of the institutional Ethics Board (REB #2335). All patients were diagnosed with invasive breast cancer and had undergone NAT prior to either mastectomy or breast conservative surgery with either sentinel or axillary lymph node excision between 2015 and 2018. Patients treated with NAT had chemotherapy and those with HER2 positive tumors received trastuzumab as well. None of the patients in this cohort had been treated with neoadjuvant checkpoint inhibitors or endocrine therapies. Histologic types, hormone receptor and HER2 status of the patients in this cohort are displayed in Table 1. All H&E-stained slides of resected lymph nodes were identified according to the grossing section in the pathology report. All slides were collected and reviewed to record the lymph node status based on any significant findings (either tumor deposit or post NAT tissue alteration with or without residual tumor). In clinical practice, immunohistochemical (IHC) stains (cytokeratins) were used to identify tumor cells that were difficult to characterize on routine H&E slides especially in cases with ITCs. However, the Post-NAT-LN dataset only contains routine H&E slides, and thus all algorithms trained in our study *did not* see IHC stains.

All lymph node slides with positive findings were scanned in the Department of Diagnostic and Molecular Pathology at 40× magnification (0.25µm/pixel) using an Aperio AT Turbo 1757 scanner (Leica Biosystems Inc., Buffalo Grove, Illinois). At least one representative slide with a negative lymph node per patient was scanned when available. In the case of multiple levels, only one representative slide was scanned. The dataset is similar to the size of CAMELYON, consisting of 293 patients and 1027 WSIs, split on the patient-level with 905 slides (494 positive, 411 negative) in the training set, and 122 (67 positive, 55 negative) in the testing set. Over all lymph nodes present in the training set, 126 contain micrometastases, 410 contain macrometastases, and 83 contain ITCs. On the testing set, 16 contain micrometastases, 49 contain macrometastases, and 13 contain ITCs. The WSI dataset was annotated by a surgical pathology fellow (DB) under the supervision of an experienced breast pathologist (SN-M) using the Sedeen Viewer [22].

Only one tissue section per patient was non-exhaustively annotated for metastatic deposits. Macrometastasis was defined as a tumor greater than 2mm (annotated using a red polygon), micrometastasis was defined as a tumor between 0.2 and 2mm (annotated using a dark green polygon), ITCs were defined as tumors < 0.2mm or < 200 scattered tumor cells (annotated using a light green polygon). According to the treatment effect, in the case of a complete pathologic response in the lymph node, the tumor bed was identified by the presence of oedematous stroma with inflammatory cells, histiocytic infiltration, and stromal fibrosis without any viable tumor cells (annotated using a blue polygon or rectangular box). In the case of a negative lymph node with no evidence of a tumor deposit or treatment

response, a yellow rectangular bounding box was drawn around one representative half in the case of one bisected lymph node (size > 5mm) or on each negative lymph node in the case of multiple, tiny negative lymph nodes on one slide (< 5mm).

All slides were also annotated on the slide-level for being positive (contains either macrometastasis, micrometastasis, or ITCs) or negative (no tumor deposit or complete pathologic response after NAT). Lymph node images that only contain ITCs were labeled as positive cases due to their potential diagnostic relevance in post-NAT conditions.

The training set was rescanned at a reduced magnification of 20x (0.46µm/pixel) using a Hamamatsu scanner in order to assess the generalizability of our method to cross-scanner domain shift. The rescanned test dataset was reduced from 122 WSIs to 115 due to an inability to find or properly rescan 7 slides.

### MSK

A subset of the test dataset used in Campanella et al. [10] was publicly released with *axillary* lymph nodes resected from a mix of chemo-naive and post-NAT patients. Slides were scanned at a lower resolution of 20x magnification (0.50 µm/pixel) using Leica Biosystems AT2 digital slide scanners. We used this dataset as a completely unseen and out-of-distribution test set to evaluate our networks. The proportion of patients that had and had not received NAT was not publicly released. The dataset consists of 78 patients and 130 WSIs (36 positive, 94 negative). The WSIs were annotated on the slide-level for being positive or negative for cancer. This dataset does not provide detailed annotations for metastatic deposits, and potentially contains slides that only contain ITCs. These purely ITC cases were also labeled as positive.

### Labeled Patch Extraction

Using the detailed annotations provided in the Post-NAT-LN and CAMELYON datasets, we extracted labeled patches to explicitly train the feature extractor. Due to the annotations differing between datasets, we adjusted the extraction strategy based on the dataset. Firstly, we created a foreground tissue mask by thresholding the image saturation greater than 1.1x the mean saturation, effectively selecting regions in the image containing color. All patches were extracted at 20x magnification (0.5 microns per pixel) and with a size of 224x224 pixels. We ensured that the test data is not used in the training stage and that the stratification is performed at the patient level.

In all except 13 slides, the CAMELYON16 dataset was exhaustively annotated for metastases by pathologists. In these fully annotated slides, positive patches were extracted which overlap with >50% of pathologist positive annotations with a stride of 56px (0.25 times patch size). In exhaustively annotated *positive* slides, we also extracted *negative* patches that have 100% overlap with negative annotations, or 0% overlap with positive annotations, all with a stride of 448px. In negative slides which do not contain annotations, negative patches were also extracted with a stride of 448px. From the non-exhaustively annotated CAMELYON17 dataset, we extracted positive patches with >50% overlap with positive annotations at a stride of 56px. We also extracted negative patches with 100% overlap with negative annotations. We then extracted negative patches from negative slides using a

stride of 448px. The CAMELYON16 and CAMELYON17 patches were combined to yield a general “CAMELYON” dataset.

The internal Post-NAT-LN dataset is also not exhaustively annotated. Thus, to obtain positive patches from positive slides, we extracted patches with >20% overlap with positive annotations (red, dark green, or light green) at a stride of 112px. We only extracted *negative* patches from *positive* slides if they had 100% overlap with the provided negative annotations at a stride of 112px. The dataset also contains annotations for full lymph nodes that are negative (yellow bounding box) and we extracted negative patches from these regions at 100% overlap with annotations and 112px stride, regardless of whether the full slide was positive or negative. Finally, we used annotations of treatment effect regions attributed to NAT (blue polygon) to extract negative patches containing treatment effects with 100% annotation overlap at 112px stride.

The stride of patch extraction and overlap with pathologist annotations were chosen depending on the dataset to ensure that the final patch datasets, CAMELYON and Post-NAT-LN, are approximately the same size and balanced between positive and negative patches. The overall sizes of patch datasets are shown in Table 2. Train and test splits were generated on the patient-level. To avoid data leakage of the test slides, the patches used to train and evaluate feature extractors were drawn solely from slides found in the slide-level *training* datasets.

## Data Splits

To evaluate the relevance of including post-NAT treatment effect features in training datasets, we created two data splits for both the WSI-level classification task and the patch-level classification task. For both tasks, the first split consists of only the CAMELYON dataset. This split contains only chemo-naive patients, and hence, we refer to this data split as the “chemo-naive” split. The second data split consists of both the CAMELYON and Post-NAT-LN datasets, and forms the “post-NAT” split. We can assess differences in performance between networks trained on these two splits to investigate the importance of post-NAT data for both the patch-level task and the slide-level task.

Concretely, we investigated performance in four scenarios:

A. **Patch features from chemo-naive feature extractor used to train the chemo-naive MIL network.** This scenario represents the results of training and evaluating our full pipeline on chemo-naive patients (i.e., developed entirely with CAMELYON datasets). Relative to the post-NAT test set, this is trained using out-of-distribution (OOD) data on both slide- and patch-levels.
B. **Patch features from chemo-naive feature extractor used to train the post-NAT MIL network.** This scenario could be practically encountered when slide-level labels are available from post-NAT patients, but not detailed annotations. The feature extractors are first trained using the chemo-naive patch dataset. Then, the trained extractor is used to extract patch features to train MIL methods on the post-NAT data split. Relative to the post-NAT test set, this train set is OOD only on the patch-level.
C. **Patch features from post-NAT feature extractor used to train the chemo-naive MIL network.** This scenario could arise when a feature extractor trained on data from a separate institution is used to extract features to train an MIL algorithm on an

internally labeled dataset of chemo-naive patients (e.g., if another institution with a labeled dataset applied our pretrained feature extractor). Relative to the post-NAT test set, this train data is OOD only on the slide-level.

D. **Patch features from post-NAT feature extractor used to train the post-NAT MIL network.** This scenario occurs when both slide-level and patch-level annotations are available for post-NAT patients. Relative to the post-NAT test set, this data split is in-distribution for both slide- and patch-level tasks.

## Feature Extractor

Feature extractors are DL networks that summarize key discriminative information inside a patch as a multidimensional set of numbers known as a feature vector. The most common networks used are Convolutional Neural Networks (CNNs) that are typically trained on natural images [23]. For this study, we take advantage of the detailed annotations that are available in both CAMELYON and Post-NAT-LN datasets to train highly clinically relevant feature extractors.

### Training

Supervised patch-level feature extractors were trained using the extracted patch dataset splits on a simple patch-level classification task. The chemo-naive feature extractor was trained on the 'tumor' versus 'no tumor' binary classification task. The post-NAT feature extractor was trained to perform three-class classification between "tumor," "negative," and "treatment effect" classes.

For each data split, five feature extractors were trained by performing 5-fold cross validation on training patches, with train/validation patches split on patient-level. The feature extractors are ResNet50 [24] CNNs that were pre-trained using self-supervised contrastive learning on histopathology data [25]. During training, image patches were randomly augmented using random vertical and horizontal flipping, random 90 degree rotations, jittering in Haematoxylin-Eosin-DAB (HED) colorspace[1], and color jittering. We also performed initial experiments without HED jittering and pre-training, which greatly reduced performance and robustness to artifacts. Further information on these experiments can be found in Supplementary Tables 5-9. Finally, due to excessive class imbalance between positive and negative classes relative to the treatment effect class, the post-NAT feature extractor dataset was balanced when training each cross validation fold by randomly duplicating patches belonging to the treatment effect class.

### Inference

After training, patch-level features were separately extracted for the chemo-naive split and post-NAT split using all five feature extractors per split. Features were extracted for all slide datasets (Post-NAT-LN, CAMELYON, MSK). Patches used to extract features have similar magnification (20x) and size (224px) as the labeled patch datasets. The stride for extracting patches for feature extraction is 224px (no overlap).

[1] We would like to acknowledge Jonathan Mazurski for the original implementation of this module.

### Heatmap Creation

We created interpretable heatmaps using the trained feature extractors as an auxiliary output to our pipeline. These heatmaps were generated by averaging the tumor probability output from each of the five feature extractors trained per data split. Heatmap creation occurs jointly with extracting patch-level features and consequently, does not greatly slow the overall pipeline. For the post-NAT data split, we were also able to generate “treatment effect” heatmaps using the predicted probability that patches belong to the treatment effect class.

## MIL Aggregators

In this work, we investigated the usage of five common MIL aggregation techniques trained on top of extracted supervised features. To combine information from all five feature extractors trained on cross validation folds, we simply concatenated the feature vectors before performing aggregation.

The first two more traditional MIL aggregation methods are max-pooling on the patch-level tumor probabilities, and max-pooling on the extracted patch features. In previous literature, these methods have been described as “instance-level” max-pooling and “embedding-level” max-pooling respectively [26]. Instance-level max-pool involves taking the highest patch probability output by the feature extractor and using it as the slide prediction. This simulates how previous works perform slide-level classification using *supervised* feature extractors. Contrarily, embedding-level max-pool passes all extracted patch-level feature vectors through an additional linear network, taking the highest probability instance as the slide-level probability. This extracted slide-level probability is compared against the slide label, and the training signal is propagated to update the weights of the linear network.

We also investigated attention-based MIL pooling [26]. This method looks to train the MIL pooling method more explicitly by learning the relative importance of the extracted patch features. Concretely, the attention aggregation method is performed using a weighted average of all the patch features. The weights, or attention placed on each patch, are computed through training a fully-connected network. The weighted average feature vector is passed through a final linear layer to obtain the slide-level probability. Gated attention was proposed alongside standard attention to improve expressivity of the learned patch importances. This method adds an additional gating mechanism to the original attention network.

The final method we explored is Clustering-constrained Attention Multiple Instance Learning (CLAM) [18]. This method further refines attention-based methods by using the learned attention values to cluster patches highly relevant to the slide-level classification and separate them from irrelevant patches.

All MIL models were trained on the weakly-labeled slide-level ‘tumor’ versus ‘no tumor’ binary classification task. Four separate sets of MIL networks were trained for each Scenario A-D (Methods and Materials - Data Splits). For each Scenario, five MIL models were trained by performing 5-fold cross validation on training WSIs, with the training and validation sets split on the patient-level. The final ensemble slide-level prediction was obtained by averaging the prediction probability output from the MIL model trained on each fold.

Notably, while using multiple 5-fold cross validation stages was more costly to train, model ensembling is known to improve robustness [27]. Furthermore, performing inference was fast when implemented efficiently, with the full pipeline taking approximately 3 minutes per slide.

## Implementation details

Our code was implemented using PyTorch[2], and implementations of MIL aggregators were adapted from the original CLAM repository[3]. Feature extractors were ResNet50 models pretrained using self-supervised contrastive methods[4] . The trained model with the highest performance on validation sets (in both MIL aggregator and feature extractor training) was evaluated on the test set to obtain final reported performance.

For supervised feature extractor training, we used the AdamW [28] optimizer with a learning rate of 0.001, weight decay of 0.01, and CrossEntropy loss. Accounting for the size of each labeled patch dataset, we trained for 20 epochs with the chemo-naive feature extractor, and 6 epochs for the post-NAT feature extractor (as it is approximately 3x larger). In both cases, the training time was 12 hours per fold with a batch size of 256 on a single A100-SXM4-80GB GPU.

For weakly-supervised MIL aggregator training, we used the Adam [29] optimizer with a learning rate of 0.0002, weight decay of 1e-5, and CrossEntropy loss. We trained aggregators for 50 epochs, taking approximately 3 hours per fold on a single A100-SXM4-80GB GPU.

# Experiments and Results

We performed experiments to evaluate feature extractors trained on chemo-naive and post-NAT splits, and MIL aggregators trained for Scenarios A-D. Results will be reported on the CAMELYON, Post-NAT-LN, MSK, and rescanned Post-NAT-LN test sets. Since exhaustive detailed annotations were not available for Post-NAT-LN and MSK, we investigated feature extractor performance using instance-level max-pool results as surrogate measures. These experiments did not train an aggregator, but directly performed max-pool on patch-level probabilities instead. Thus, they were more directly related to feature extractor quality. MIL aggregators trained in Scenarios A-D were assessed on the slide-level binary classification task on all test datasets. Broadly, performance was assessed on slide-level tasks via accuracy and AUC metrics on predicted slide-level probabilities.

5-fold cross validation splits (for slide-level training and patch-level training) were kept consistent across experiments. Thus, the reported statistical significance was computed by using an unpaired nonparametric Mann-Whitney U test on the metrics obtained when independently evaluating models trained per fold on test sets. Final reported results come from ensembling models trained in each fold.

[2] https://pytorch.org/
[3] https://github.com/mahmoodlab/CLAM
[4] https://github.com/ozanciga/self-supervised-histopathology

## Instance-Level Max-pool and Interpretable Heatmaps

Instance-level max-pooling is highly susceptible to false positive predictions. Therefore, we report two values for accuracy; first using a 0.5 probability threshold (i.e., positive slide probability >0.5 and vice versa for a negative slide) and then using an “adaptive” threshold. To obtain the adaptive threshold, we varied the threshold value for each test set and selected the value that gave the highest accuracy. The former threshold could be considered more “fair,” as it maintains the same probability threshold used to originally train the network, while the latter provides a better picture of the overall capabilities of the network in distinguishing positive and negative slides. Table 3 reports the results of instance-level max-pool aggregation on the slide-level classification task. Figures 2 and 3 display exemplar tumor probability heatmaps output by feature extraction using chemo-naive and post-NAT feature extractors. We additionally visualized probability maps of *treatment effect* regions for the post-NAT feature extractor.

It is difficult to fully quantify heatmap performance on the internal dataset, as slides were not exhaustively annotated. Thus, instance-level max-pool results–which are more directly related to heatmap quality–were used as a surrogate measure to quantitatively assess heatmap quality. While the chemo-naive feature extractor obtained better results on the chemo-naive test set, we found the post-NAT feature extractor performed better on all of the test sets that included post-NAT patients. Additionally, on the Post-NAT-LN test set, we found that when using a 0.5 classification threshold, the chemo-naive feature extractor classified every slide as positive (0.549 accuracy; Table 3), indicating that major calibration errors were introduced via the max-pool aggregation. The post-NAT feature extractor performed much better on this surrogate measure, obtaining a significantly better accuracy (0.713 accuracy, $p<0.01$).

Figure 2 shows a qualitative comparison between heatmaps generated from chemo-naive feature extractors versus post-NAT feature extractors on a slide belonging to the Post-NAT-LN test set. The outlined regions containing post-NAT treatment effects confounded the chemo-naive network, resulting in a large number of false positive predictions. Post-NAT feature extractor heatmaps also appear qualitatively better in treatment effect regions. Figure 3a-c displays the heatmaps for the false negative slide-level prediction containing a micrometastasis. The post-NAT feature extractor heatmap contains a single red (highly suspicious) patch directly localized on the micrometastasis missed by the slide-level prediction. The chemo-naive feature extractor heatmap also predicts the micrometastasis as suspicious, but also severely overpredicts the surrounding regions containing treatment effect as positive. This overcrowded heatmap is likely to detract from pathologist interpretability, though future work will need to assess this in a more quantitative manner.

## Embedding-Level MIL

We summarize results by reporting the best embedding-level MIL method per Scenario based on AUC. The full results separated by MIL method can be found in Supplementary Tables 1-4. Again, to maintain a fair assessment, the probability threshold remains at 0.5 for computing slide-level accuracy. We additionally report the specificity of the network when calibrating the prediction threshold to achieve 100% sensitivity (“$Spec_{100}$”). This metric has

been used in previous works to investigate the number of slides that could potentially be excluded from pathologist assessment in clinical workflows [10].

## In-Domain Test Sets

We evaluate the pipeline against the (unseen, but) in-domain test sets for Post-NAT-LN and CAMELYON. Table 4 displays the results for these experiments. We found our pipeline trained with post-NAT data performed significantly better on slide-level classification compared to the pipeline trained on fully chemo-naive data for the Post-NAT-LN dataset (0.986 AUC for Scenario D vs. 0.955 AUC for Scenario A, $p<0.01$; Table 4). We note that the loss in performance is largely recuperated when we include *slides* from post-NAT patients, even when using a feature extractor trained on chemo-naive patients (0.986 AUC for Scenario B; Table 4). The pipeline trained using post-NAT patients in both feature extractor and slide-level aggregator datasets was tied for the highest AUC on the Post-NAT-LN test set, and obtained the highest accuracy (0.986 AUC and 0.943 accuracy for Scenario D; Table 4).

Additionally, we found that the pipeline that was fully trained on the chemo-naive CAMELYON dataset obtained 0.994 AUC (Scenario A; Table 4) when tested on the CAMELYON test set. This classification AUC is equivalent to the challenge-winning algorithm of the CAMELYON16 challenge, which used feature engineering on probability heatmaps to yield a slide-level classification [2,30].

When performing error analysis on the improperly classified WSIs in the Post-NAT-LN test set using the fully post-NAT pipeline (Scenario D), we found that 5 false negative predictions contained only ITCs (out of a total of 12 in the test set), and one false negative contained a micrometastasis. This demonstrates the high level of difficulty associated with catching ITCs in patch- or slide-level outputs. The inclusion of ITCs in the Post-NAT-LN test set stands as an advantage of our dataset, especially in the post-NAT setting where they may be more relevant [14]. When we removed slides with only ITCs from the test set in order to compare our results more directly with the reported results from the CAMELYON16 challenge, we obtained a classification accuracy of 0.982 (one false positive and one false negative).

Additional analysis on the seven improperly classified WSIs from the pipeline trained in Scenario D reveals that three WSIs were taken from a ER+/HER2- patient, two came from the same ER-/HER2+ patient, and two came from triple negative patients. Since the full Post-NAT-LN testing cohort only contained four triple negative patients (seven triple negative WSIs), it is possible that the pipeline may be susceptible to misclassifying patients from this biological subtype. However, the misclassified slides were very difficult cases, with one containing a small micrometastasis and the other only containing ITCs. Thus, further work may be required to determine if this susceptibility exists or if the pipeline's incorrect classification was due to tumor size.

## Out-of-Domain Test Sets

As an evaluation of the pipeline against cross-scanner imaging protocols and cross-institutional shift, we evaluate the overall pipeline against the Post-NAT-LN test set, rescanned using a Hamamatsu scanner, and the MSK test set. Both datasets are scanned at 20x magnification, meaning the rescanned Post-NAT-LN test set constitutes a scanner

domain shift relative to training data, and MSK constitutes a scanner, patient, and institution shift. The rescanned Post-NAT-LN dataset was reduced from 122 WSIs to 115 due to an inability to find or properly rescan 7 slides. Supplementary Figure 1 shows examples of rescanned regions and corresponding heatmaps. Table 5 displays the results of these experiments.

When the full pipeline was trained using post-NAT slides, we found it reasonably robust in all metrics on the rescanned Post-NAT-LN test set (1.01% drop in AUC, 0.42% drop in accuracy, 12.13% drop in $Spec_{100}$; Table 5). Conversely, the networks trained on purely chemo-naive slides had more severe degradations in AUC (2.93% drop) and $Spec_{100}$ (90.5% drop). The network fully trained using post-NAT data performed the best in all metrics on the rescanned dataset (0.976 AUC, 0.939 accuracy, 0.623 $Spec_{100}$ for Scenario D; Table 5). This demonstrates that, when trained on post-NAT data, the overall pipeline is robust to changes in scanner type and shifts in magnification (from 40x to 20x).

On the MSK test set, we found a small decrease in AUC for the pipeline trained with post-NAT data compared to the pipeline trained with only chemo-naive data (0.921 AUC for Scenario D vs. 0.938 AUC for Scenario A, $p<0.05$; Table 5). This was coupled with large improvements to classification accuracy (0.938 accuracy vs. 0.838 accuracy, $p<0.05$; Table 5). Having the pipeline trained with post-NAT data in the feature extractor or for the slide-level aggregator (Scenarios B, C, D) broadly improved the balance between classification AUC and accuracy.

## Interpretability Study

We performed a simple experiment to evaluate the benefits of the pathologist-interpretable heatmap as an auxiliary model output. We provided a pathologist (DB) with a slide-level prediction (based on a 0.5 prediction threshold) and the patches in the slide with the top 10 highest heatmap probabilities. We asked the pathologist if they agreed or disagreed with the slide-level prediction based on the contents of the top 10 most suspicious patches. If they disagreed with the prediction, we also asked them to describe why. We provided the pathologist with the slides that were incorrectly classified by the embedding-level MIL network from Scenario D (fully post-NAT), along with 3 true positive and 3 true negative predictions by the network in the Post-NAT-LN test set. As mentioned, this network misclassified six positive slides, five of which contained only ITCs and one containing a micrometastasis, and it also misclassified a negative slide. We selected this network because it provided the best tradeoff between heatmap quality and slide-level prediction performance (Table 4; Figures 2, 3).

Figure 4a shows an example of this workflow for a false negative slide-level prediction, which contained a small micrometastasis that was successfully detected by the pathologist. Figure 4b shows an example for a true positive prediction on a slide containing only a single ITC, which was also detected by the pathologist. Figure 4c shows an example where the only false positive was corrected due to the pathologist detecting that the model was confounded by skin and benign sweat glands. Overall, the pathologist was able to correctly identify four out of six false negative predictions and as well as the single false positive prediction. The only false negative slide containing a micrometastasis (Figure 4a) was corrected by the pathologist, meaning that when removing ITCs, the full interpretable

workflow with a pathologist viewing the top 10 heatmap regions *did not misclassify a single slide in the Post-NAT-LN test set.* They also correctly identified the three true positives and negatives in all cases but one negative case where they were unsure and required additional context. Notably, DB was also involved with the original annotation process for these slides, though there was a multi-year washout period to ensure minimal bias.

# Discussion

In this study, we created Post-NAT-LN, a large WSI dataset consisting of post-NAT breast cancer patients and performed the first explicit evaluation of whether DL methods trained solely on chemo-naive patients are generalizable to this patient group. We additionally proposed a two-stage pipeline to address the question of whether feature extractors *supervised* on a highly related task can be used in conjunction with weakly labeled MIL methods. We assessed whether this simple amendment could create a pipeline that is both highly sensitive on patch-level, and able to learn the best aggregation method on the patch-level based on the training dataset.

Previous studies performing MIL that only require “weak” slide-level labels have been proposed [18,19,21,26]. When applied to histopathology, these studies generally use feature extractors that are trained to classify natural images [18,19], or extractors that are trained using self-supervised methods [20,21]. While these feature extractors have the key benefit of not requiring costly detailed annotations from pathologists, we found that efforts to extract sensitive and interpretable heatmaps are met with variable success. Specifically, the heatmaps output from these weakly supervised networks have no guarantee of having high sensitivity or detection capability, especially under domain shift and for smaller datasets, which limits their potential for clinical adoption.

There have also been methods that instead propose to use dense labels and fully supervised feature extractors to perform cancer detection tasks [11,12,30]. These studies use detailed patch-level labels to train networks which are expensive to create. However, these methods remain the gold standard for metastasis detection [20]. Manual feature engineering is typically used to obtain slide-level predictions from a fully trained feature extractor. For example, [11] uses the presence of > 2 connected tumor predicted patches to determine the slide-level class, but this approach makes it impossible to detect very small micrometastases or isolated tumor cells. Thus, rather than performing manual feature engineering, we used our proposed pipeline to investigate whether a simple data-driven approach can be used to aggregate patch-level features to a slide-level prediction.

**The overall pipeline achieves strong, robust results for metastasis detection.** We found that our proposed pipeline using MIL methods and *supervised feature extractors* performed equally well compared to methods that use *feature engineering* on patch-level heatmaps. This can be seen from the results of testing our pipeline trained with purely chemo-naive data (Scenario A) on the CAMELYON test set. Our pipeline was able to recreate the performance of the winning algorithm of the CAMELYON16 challenge which used feature engineering to obtain its slide-level classification [2,30]. We found that our pipeline also achieved generally strong classification performance on in-domain test data, while being robust to large cross-scanner and cross-institutional domain shifts.

**Post-NAT data benefits slide-level prediction calibration and sensitivity.** We found that including post-NAT data in training the feature extractor and slide-level MIL network was beneficial to performance when applied to test sets containing post-NAT patients (MSK, Post-NAT-LN). Our pipeline trained with fully chemo-naive data (Scenario A) had significantly better AUC but significantly worse accuracy than our network trained with post-NAT data (Scenario D) on the MSK test set. This may imply that, in the absence of a predefined threshold, the fully chemo-naive network was better at sorting between positive and negative slides (high AUC), but the overall prediction distribution was shifted (low accuracy). We also note that, regardless of which feature extractor was used, introducing post-NAT slides to train MIL aggregators broadly improved specificity of the pipeline at 100% sensitivity for test sets containing post-NAT patients. In other words, $Spec_{100}$ is higher for both Scenario B relative to A and for D relative to C.

Specifically on the internal Post-NAT-LN test set, we found training a slide-level MIL aggregator with post-NAT slides more important than training a patch-level feature extractor with post-NAT patches. The improvement gained from introducing post-NAT slides (Scenario B vs. A) was greater than when introducing post-NAT patches (Scenario C vs. A). We found that this discrepancy remained even after evaluating on the rescanned dataset. The ability for a MIL model to generalize despite being built on a feature extractor trained on OOD data is well documented in previous works that use feature extractors trained on natural images [18]. Given the lower amount of labeling effort necessary to produce a slide-level label, future works specifically interested in *slide-level classification* may find relatively better return on investment from obtaining more weakly labeled slides than annotating more patches.

**Using post-NAT data greatly improves the quality of interpretable heatmaps.** We found through qualitative assessment of Figures 2 and 3 along with the surrogate quantitative measures from instance-level max-pool results in Table 3 that post-NAT data improves the quality of interpretable heatmaps. The improvement of heatmap quality from the post-NAT feature extractor could potentially be attributed to the introduction of the *treatment effect* class. Figures 2d and 2h display example heatmaps generated by the post-NAT feature extractor. For the slide from the Post-NAT-LN test set, high values in the treatment effect heatmap are highly localized to the pathologist annotation (in cyan). For the slide from the chemo-naive CAMELYON test set, the treatment effect heatmap is completely empty. This could indicate that the introduction of the treatment effect class is driving the improvement of specificity in post-NAT feature extractor heatmaps. The generation of a heatmap showing treatment effects may also be a useful addition to the workflow as there is some evidence that the detection of treatment effect in axillary lymph nodes after neoadjuvant chemotherapy identifies a subset of patients with an outcome intermediate between that of completely node-negative and node-positive patients [31].

Results on instance-level max-pool were reported using both “fair” thresholds (0.5) and “best” thresholds (chosen to maximize classification accuracy on the test set) because determining new classification thresholds to suit the test dataset is a nontrivial task, and an active research area [32]. We leave the assessment of the best threshold to future work on deploying the algorithm. One example way to perform this type of model “calibration” is to simply set the decision threshold based on model predictions on a small labeled set of local deployment data.

**High quality heatmaps benefit interpretability and expedite error correction on the slide-level.** Our initial study on the benefits of interpretability is a simple example for how an accurate heatmap may improve pathologist workflows. Our results show that the pathologist was quickly able to detect errors in slide-level classification using the heatmap output from the post-NAT feature extractor. Further studies will be needed to determine the efficacy of our interpretable pipeline in a clinical context, but these initial results suggest that accurate heatmaps make it possible for pathologists to assess the reliability of the slide-level prediction.

One limitation of this study is the exploratory nature of our experiments on interpretability. In a clinical assessment, it is unreasonable for a pathologist to only view isolated patches. Thus, future work may instead integrate this workflow into a WSI viewing software by providing the top 10 *patch locations* that have the highest predicted tumor class probability in the slide. This would allow a pathologist to obtain more surrounding context to inform their assessment. The true benefits of the overall pipeline cannot be discerned without bringing the pathologist into the loop. In other words, future work must involve *quantifying* the benefits of this algorithm in user studies to determine the importance of slide-level classification versus heatmap quality. By doing so, we could better probe if the interpretable algorithm reduces pathologist errors, speeds up workflows, and potentially increases the acceptability of the algorithm.

In this work, we solely investigated breast cancer patients. Although our data included annotations for metastatic carcinoma and examples of altered histology due to NAT, we did not explicitly annotate cases with other findings such as benign epithelial inclusions, non-mammary epithelial, or other malignancies and non-neoplastic, clinically significant findings (i.e. granulomas). This work investigated lymph node histology from NAT with the aim of creating a simple decision and guidance tool for pathologists. Future work may point toward enriching the pipeline with the ability to detect these potentially relevant findings.

Another limitation of our dataset formulation is that cases with invasive lobular carcinoma (ILC) are underrepresented. In the CAMELYON dataset, only 12% of the cases were patients with ILC. In our institution, patients with ILC rarely receive NAT, hence Post-NAT-LN did not contain any ILC cases. Jarkman et al. [12] showed that a model trained on CAMELYON data struggled to detect ILC in a second dataset (AIDA [33]). We similarly found that the network struggled with ILC cases in the MSK dataset, and particularly in cases with low nuclear grade. These difficult cases could be associated with the relatively low $Spec_{100}$ values obtained on the MSK test set. Future work may need to incorporate greater numbers of ILC cases to improve performance on this patient subgroup.

The full, proposed method could be added to clinical workflows by performing inference on incoming patients overnight and having slide-level predictions and heatmaps ready for pathologists to view in the morning. The slide-level tumor probability output can help pathologists with prioritization, guiding focus to complex slides, or simply act as a second opinion to ensure slides are not improperly classified. The interpretable heatmap can guide pathologists to key regions of interest in a WSI, speed up the overall workflow, reduce false negative predictions for micrometastasis or ITCs, and promote confidence in the algorithm. Overall, this work has the potential to greatly benefit routine metastatic assessment in lymph

nodes by streamlining workflows, freeing pathologists to focus on other critical tasks and mitigating false negative predictions.

## Ethics Approval

Ethics approval for this study was obtained from the institutional Ethics Board of Sunnybrook Health Sciences Centre, Toronto, Canada (REB #2335). All patient slides are deidentified and retrospectively collected for this study.

## Author Contribution Statement

T.X. contributed to writing, method formulation, model training, inference and data analysis. D.B. contributed to writing, data collection and annotation, and data analysis. C.S. and M.S.W.L. contributed to method formulation, data processing, and initial stage network training and inference. M.G. supervised the project. S.N. contributed to project conception, writing, data collection and analysis, and provided project supervision. A.L.M. contributed to project conception, writing, data collection, and provided project supervision. All authors read and approved the final paper.

## Data Availability Statement

The Post-NAT-LN dataset cannot be released due to institutional restrictions. The CAMELYON dataset can be found on the challenge webpage (https://camelyon17.grand-challenge.org/Data/), and the MSK test set can be found in the original work [10].

## Funding Statement

This work was funded by Canadian Cancer Society (grant #705772) and Canadian Institutes of Health Research (CIHR grant #162327). This work was additionally supported by funding from the Canada Foundation for Innovation (40206), and the Ontario Research Fund. T.X. is partially funded by the NSERC PGS-D award. A.L.M. is partially supported by the Tory Family Chair in Oncology.

# References


[1] Huang SC, Chen CC, Lan J, et al. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nat Commun*. 2022;13(1):3347.

[2] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 2017;318(22):2199-2210.

[3] Giammanco A, Bychkov A, Schallenberg S, et al. Fast-track development and multi-institutional clinical validation of an artificial intelligence algorithm for detection of lymph node metastasis in colorectal cancer. *Mod Pathol*. Published online April 16, 2024:100496.

[4] Symmans WF, Peintinger F, Hatzis C, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol*. 2007;25:4414-4422.

[5] Symmans WF, Wei C, Gould R, et al. Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J Clin Oncol*. 2017;35(10):1049-1060.

[6] Burgués O, López-García MÁ, Pérez-Míes B, et al. The ever-evolving role of pathologists in the management of breast cancer with neoadjuvant treatment: recommendations based on the Spanish clinical experience. *Clin Transl Oncol*. 2018;20(3):382-391.

[7] Bandi P, Geessink O, Manson Q, et al. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Trans Med Imaging*. 2019;38(2):550-560.

[8] Steiner DF, MacDonald R, Liu Y, et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am J Surg Pathol*. 2018;42(12):1636-1646.

[9] Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience*. 2018;7(6). doi:10.1093/gigascience/giy065

[10] Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309.

[11] Khan A, Brouwer N, Blank A, et al. Computer-assisted diagnosis of lymph node metastases in colorectal cancers using transfer learning with an ensemble model. *Mod Pathol*. 2023;36(5):100118.

[12] Jarkman S, Karlberg M, Pocevičiūtė M, et al. Generalization of Deep Learning in Digital Pathology: Experience in Breast Cancer Metastasis Detection. *Cancers* . 2022;14(21). doi:10.3390/cancers14215424

[13] Challa B, Tahir M, Hu Y, et al. Artificial Intelligence-Aided Diagnosis of Breast Cancer Lymph Node Metastasis on Histologic Slides in a Digital Workflow. *Mod Pathol*. 2023;36(8):100216.

[14] Wong SM, Almana N, Choi J, et al. Prognostic significance of residual axillary nodal micrometastases and isolated tumor cells after neoadjuvant chemotherapy for breast cancer. *Ann Surg Oncol*. 2019;26(11):3502-3509.

[15] van der Heiden-van der Loo M, Schaapveld M, Ho VKY, Siesling S, Rutgers EJT, Peeters PHM. Outcomes of a population-based series of early breast cancer patients with micrometastases and isolated tumour cells in axillary lymph nodes. *Ann Oncol*. 2013;24(11):2794-2801.

[16] Liikanen JS, Leidenius MH, Joensuu H, Vironen JH, Meretoja TJ. Prognostic value of isolated tumour cells in sentinel lymph nodes in early-stage breast cancer: a prospective study. *Br J Cancer*. 2018;118(11):1529-1535.

[17] Houvenaeghel G, Classe JM, Garbay JR, et al. Prognostic value of isolated tumor cells and micrometastases of lymph nodes in early-stage breast cancer: a French sentinel node multicenter cohort study. *Breast*. 2014;23(5):561-566.

[18] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555-570.

[19] Shao Z, Bian H, Chen Y, et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classication. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, eds. *Neural Inf Process Syst*. 2021;34:2136-2147.

[20] Dehaene O, Camara A, Moindrot O, de Lavergne A, Courtiol P. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012 03583*. Published online 2020. http://arxiv.org/abs/2012.03583

[21] Chen RJ, Chen C, Li Y, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2022:16144-16155.

[22] Martel AL, Hosseinzadeh D, Senaras C, et al. An image analysis resource for cancer research: PIIP-pathology Image Informatics Platform for visualization, analysis, and management. *Cancer Res*. 2017;77(21):e83-e86.

[23] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2009:248-255.

[24] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[25] Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. *Mach Learn Appl*. 2022;7(100198):100198.

[26] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. Dy J, Krause A, eds. *ICML*. 2018;80:2132-2141.

[27] Petrick N, Akbar S, Cha KH, et al. SPIE-AAPM-NCI BreastPathQ challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *J Med Imaging (Bellingham)*. 2021;8(3):034501.

[28] Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv [csLG]*. Published online November 14, 2017. https://openreview.net/forum?id=Bkg6RiCqY7

[29] Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv [csLG]*. Published online December 22, 2014. http://arxiv.org/abs/1412.6980

[30] Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. *arXiv [q-bioQM]*. Published online June 18, 2016. http://arxiv.org/abs/1606.05718

[31] Newman LA, Pernick NL, Adsay V, et al. Histopathologic evidence of tumor regression in the axillary lymph nodes of patients treated with preoperative chemotherapy correlates with breast cancer outcome. *Ann Surg Oncol*. 2003;10(7):734-739.

[32] Silva Filho T, Song H, Perello-Nieto M, Santos-Rodriguez R, Kull M, Flach P. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach Learn*. 2023;112(9):3211-3260.

[33] Jarkman S, Lindvall M, Hedlund J, Treanor D, Lundstrom C, Van Der Laak J. Axillary lymph nodes in breast cancer cases. Published online November 19, 2019. doi:10.23698/AIDA/BRLN

# Figure Legends

*Figure 1: Overall workflow for methodology.*

*Figure 2: Heatmaps generated by chemo-naive and post-NAT feature extractors and overlaid on WSI regions.* ***a)-d)*** *Example from Post-NAT-LN test set showing improved specificity (reduced false positives) for the post-NAT feature extractor in treatment effect regions.* ***a)*** *Pathologist annotation overlaid on original image, green outlines micrometastases (non-exhaustively), cyan outlines the treatment effect region.* ***b)*** *Heatmap produced by chemo-naive feature extractor.* ***c)*** *Heatmap produced by post-NAT feature extractor.* ***d)*** *Heatmap output from post-NAT feature extractor for the third "treatment effect" class.* ***e)-h)*** *Example from CAMELYON test set displaying heatmap quality and proper lack of predicted treatment effect by the post-NAT model.* ***e)*** *Original slide region.* ***f)*** *Heatmap produced by chemo-naive feature extractor.* ***g)*** *Heatmap produced by post-NAT feature extractor.* ***h)*** *Heatmap of treatment effect class produced by post-NAT feature extractor.*

*Figure 3: Heatmaps generated by chemo-naive and post-NAT feature extractors and overlaid on WSI regions displaying high specificity for two cases with very minimal metastatic deposits.* ***a)-c)*** *Case containing a single micrometastasis.* ***a)*** *Original slide region.* ***b)*** *Heatmap produced by chemo-naive feature extractor.* ***c)*** *Heatmap produced by post-NAT feature extractor with zoomed in view of detected micrometastasis in a single patch.* ***d)-f)*** *Case containing ITCs.* ***d)*** *Original slide region.* ***e)*** *Heatmap produced by chemo-naive feature extractor.* ***f)*** *Heatmap produced by post-NAT feature extractor with zoomed in view of detected ITCs.*

*Figure 4: Top 10 patches based on predicted tumor probability output from the post-NAT feature extractor for pathologist interpretability study.* ***a)*** *Slide-level false negative case that was caught and corrected by the pathologist during the interpretability study (yellow arrow indicates detected micrometastasis).* ***b)*** *Slide-level true positive case that was accurately confirmed by the pathologist during the interpretability study (yellow arrows indicate detected ITCs).* ***c)*** *Slide-level false positive case containing benign skin and sweat glands that was caught and corrected by the pathologist during the interpretability study.*

| | **Total** | | | | **Train** | | | | **Test** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Hormone Receptor and HER2 Status** | | | | | | | | | | | |
| | **Pos.** | **Low Pos.** | **Neg.** | **N/A** | **Pos.** | **Low Pos.** | **Neg.** | **N/A** | **Pos.** | **Low Pos.** | **Neg.** | **N/A** |
| **ER** | 170 | 11 | 98 | 14 | 142 | 9 | 89 | 13 | 28 | 2 | 9 | 1 |
| **PR** | 89 | 39 | 149 | 16 | 77 | 31 | 131 | 14 | 12 | 8 | 18 | 2 |
| **HER2** | 108 | N/A | 168 | 17 | 92 | N/A | 146 | 15 | 16 | N/A | 22 | 2 |
| | **Patient Subtypes** | | | | | | | | | | | |
| | **ER+/ HER2 -** | **ER+/ HER2 +** | **ER-/ HER2 -** | **N/A** | **ER+/ HER2 -** | **ER+/ HER2 +** | **ER-/ HER2 -** | **N/A** | **ER+/ HER2 -** | **ER+/ HER2 +** | **ER-/ HER2 -** | **N/A** |
| | 106 | 73 | 62 | 52 | 88 | 62 | 58 | 45 | 18 | 11 | 4 | 7 |
| | **Nuclear Grade** | | | | | | | | | | | |
| | **1** | **2** | **3** | **N/A** | **1** | **2** | **3** | **N/A** | **1** | **2** | **3** | **N/A** |
| | 25 | 161 | 84 | 23 | 16 | 145 | 74 | 18 | 9 | 16 | 10 | 5 |

*Table 1: Post-NAT-LN patient cohort characteristics including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor (HER2) status, biological subtypes described by ER and HER2 status, and nuclear grade. Total number of patients and patients separated into train and test splits are shown. Tumors were graded on primary residual tumor post treatment when available. In cases where there was no residual invasive carcinoma in the breast, grade was retrieved from biopsy or assigned based on lymph node metastasis.*

| Dataset | Total patches ($N_{pos}/N_{neg}/N_{treatment}$) | | | Training patches ($N_{pos}/N_{neg}/N_{treatment}$) | | | Testing patches ($N_{pos}/N_{neg}/N_{treatment}$) | | |
|---|---|---|---|---|---|---|---|---|---|
| **CAMELYON** | 2,750,858 | | | 2,106,577 | | | 644,281 | | |
| | 1,571,062 | 1,179,796 | 0 | 1,197,515 | 909,062 | 0 | 373,547 | 270,734 | 0 |
| **Post-NAT-LN** | 2,717,106 | | | 1,893,400 | | | 823,706 | | |
| | 811,058 | 1,414,872 | 491,176 | 554,625 | 1,104,337 | 234,438 | 256,433 | 310,535 | 256,738 |

*Table 2: Labeled patch dataset extracted with number of patches in each annotation class. $N_{pos}$ = positive for tumor, $N_{neg}$ = no tumor or treatment effect, $N_{treatment}$ = post-NAT treatment effects present.*

| | **CAMEYLON Test** | | | **Post-NAT-LN Test** | | | **MSK Test** | | |
|---|---|---|---|---|---|---|---|---|---|
| Extractor Split | AUC | ACC (adapt.) | ACC (fair) | AUC | ACC (adapt.) | ACC (fair) | AUC | ACC (adapt.) | ACC (fair) |
| Chemo-naive | **0.976** | **0.969** | **0.380** | 0.947 | 0.893 | 0.549 | 0.886 | 0.885 | 0.277 |
| Post-NAT | 0.946 | 0.946 | **0.380** | **0.963** | **0.910** | **0.713** | **0.900** | **0.915** | **0.292** |

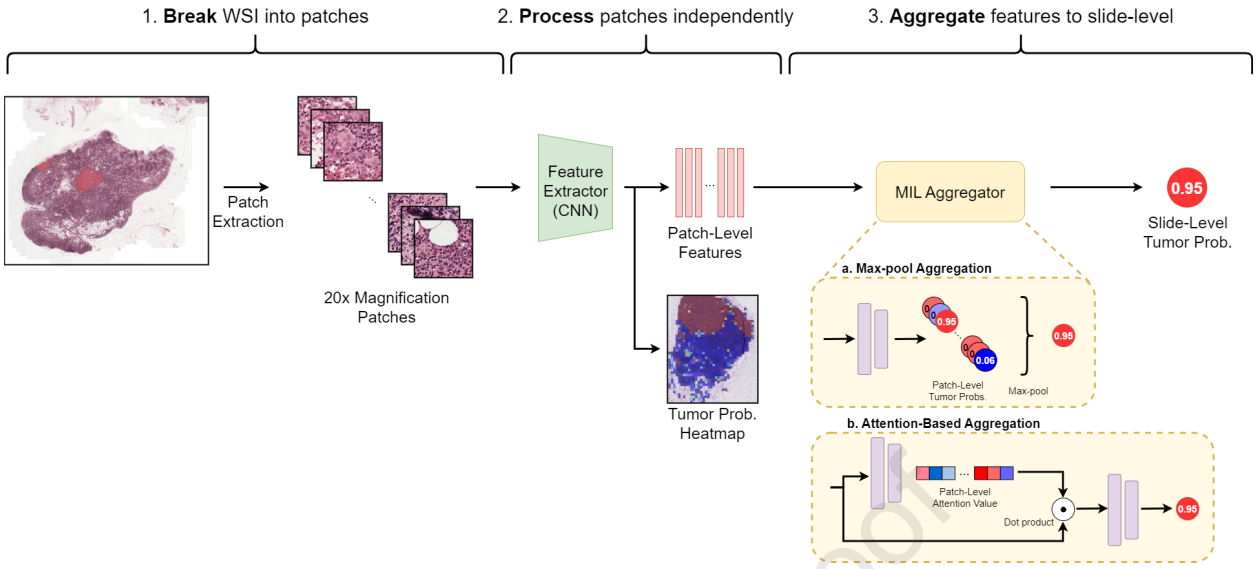
*Table 3: Instance-level max-pool AUC, accuracy with adaptive threshold (maximizes accuracy on test set), and accuracy with fair threshold (0.5) on three test sets reported for each feature extractor.*

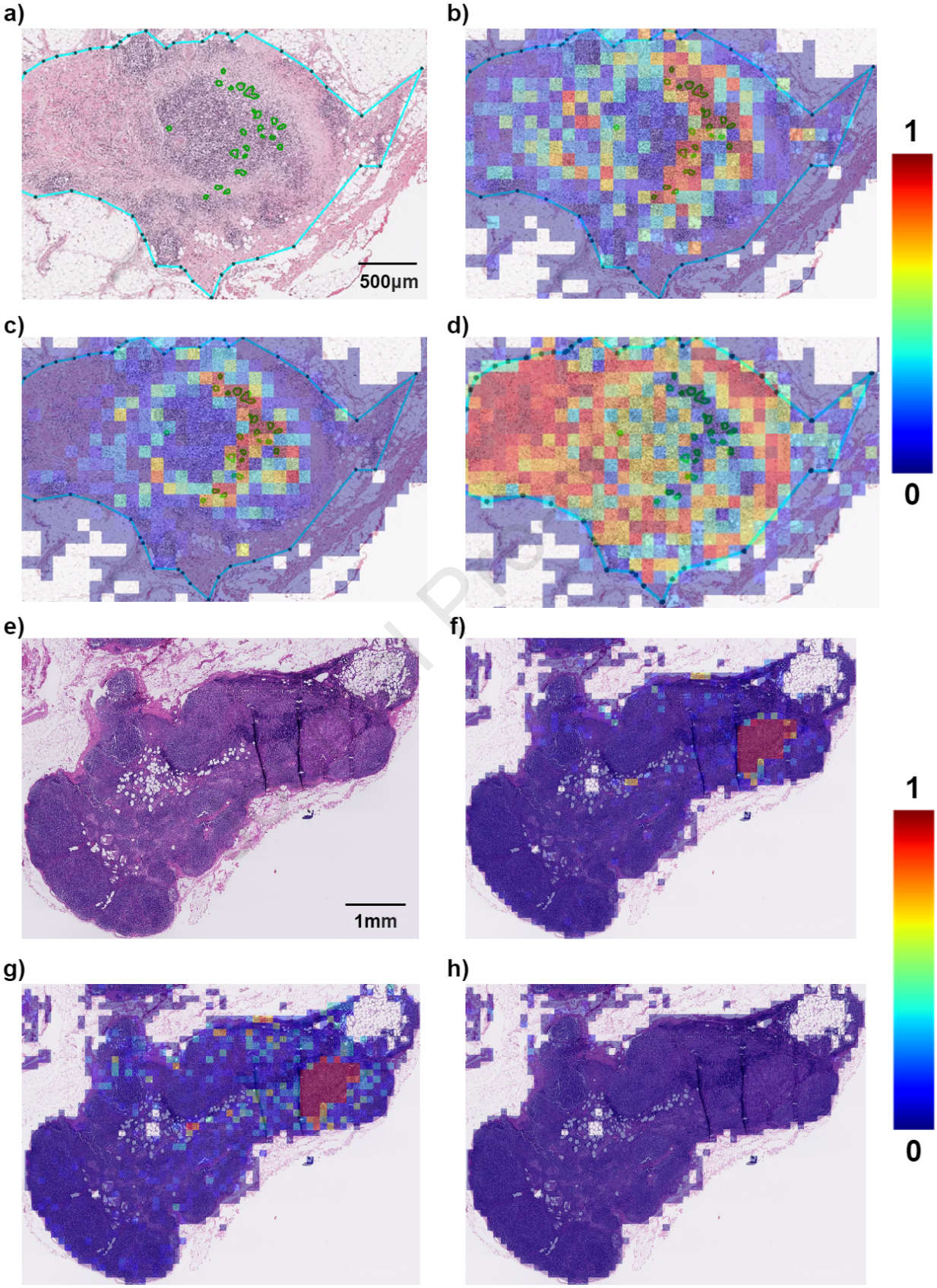
| | | | **CAMEYLON Test** | | | | **Post-NAT-LN Test** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Extractor Split | Slide Split | Best MIL | AUC | ACC | $Spec_{100}$ | Best MIL | AUC | ACC | $Spec_{100}$ |
| A | Chemo-naive | Chemo-naive | Gated | **0.994** | 0.969 | 0.838 | Max. | 0.955 | 0.910 | 0.400 |
| B | Chemo-naive | Post-NAT | CLAM | **0.994** | **0.977** | **0.850** | Max. | **0.986** | 0.926 | **0.782** |
| C | Post-NAT | Chemo-naive | Attn. | 0.992 | 0.953 | 0.838 | Attn. | 0.972 | 0.910 | 0.473 |
| D | Post-NAT | Post-NAT | CLAM | 0.988 | 0.946 | 0.75 | Gated | **0.986** | **0.943** | 0.709 |

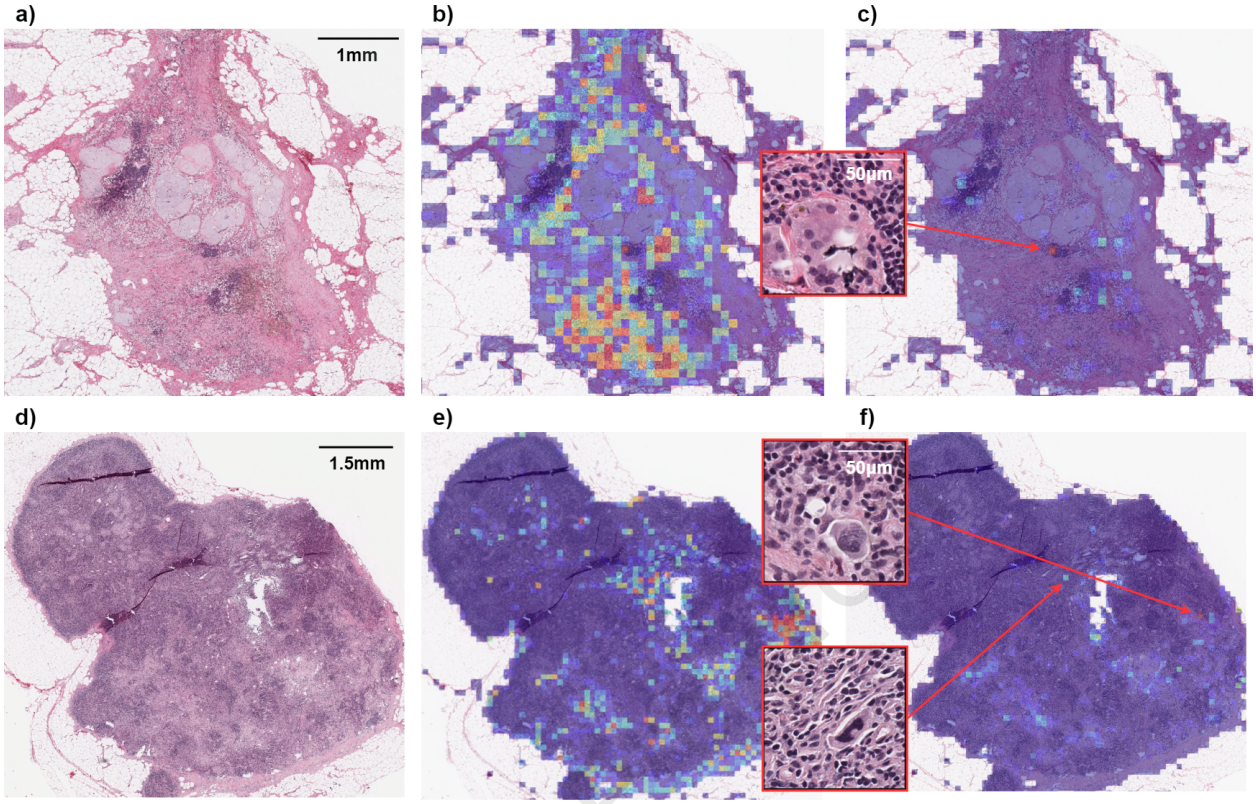
*Table 4: Embedding-level AUC, accuracy with 0.5 threshold, and specificity at 100% sensitivity ($Spec_{100}$) metrics for best MIL aggregation method for two test sets reported for Scenarios A-D. Attn. = Attention-based aggregation; Gated = Gated attention; Max. = Maxpool aggregation; ACC = Accuracy.*

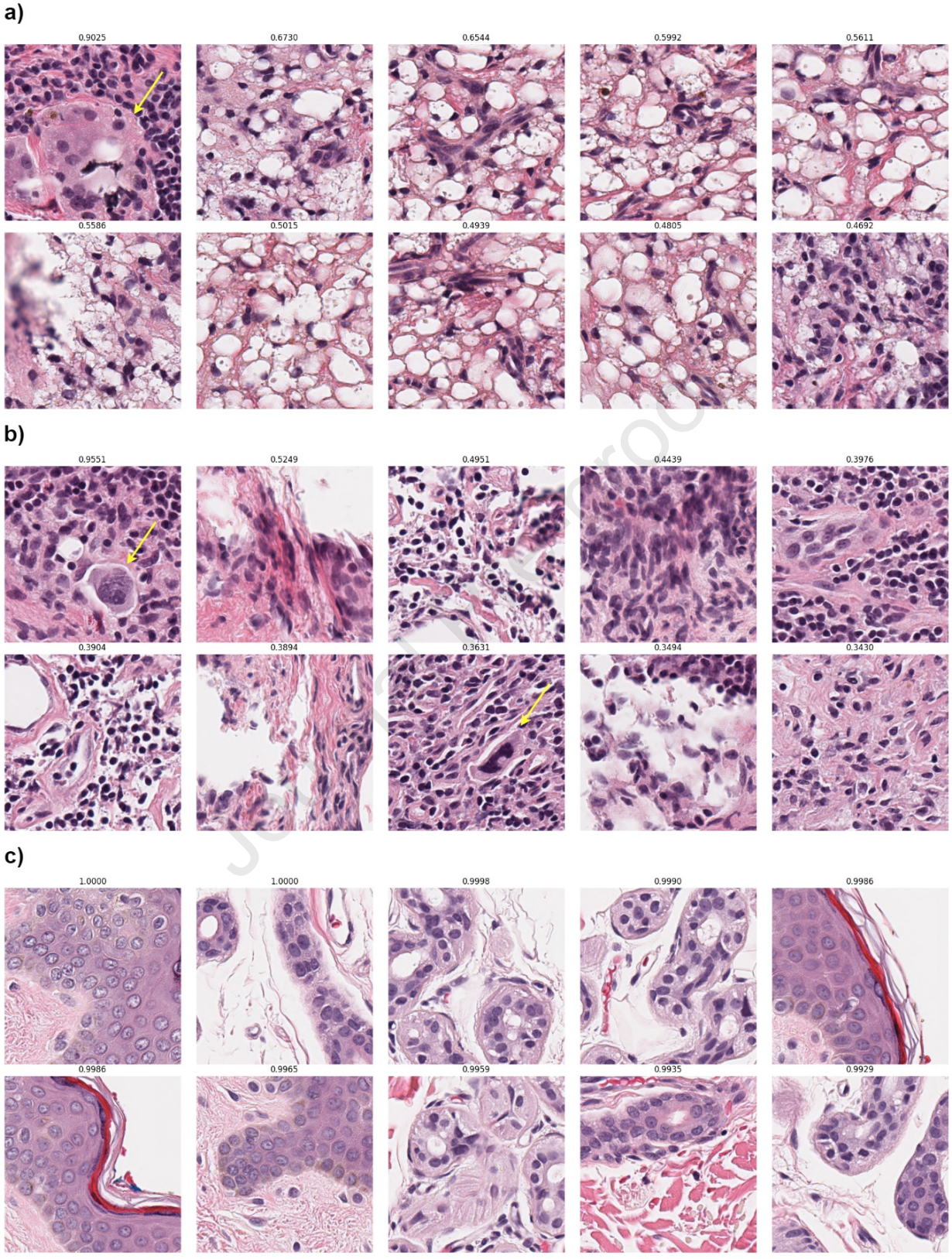
| | | | **Rescanned Post-NAT-LN Test** | | | | **MSK Test** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Extractor Split | Slide Split | Best MIL | AUC | ACC | $Spec_{100}$ | Best MIL | AUC | ACC | $Spec_{100}$ |
| A | Chemo-naive | Chemo-naive | CLAM | 0.927 | 0.922 | 0.038 | Attn. | **0.938** | 0.838 | 0.266 |
| B | Chemo-naive | Post-NAT | Gated | 0.971 | 0.930 | 0.566 | Attn. | 0.921 | 0.908 | **0.340** |
| C | Post-NAT | Chemo-naive | Attn. | 0.953 | 0.913 | 0.038 | CLAM | 0.919 | **0.946** | 0.064 |
| D | Post-NAT | Post-NAT | Gated | **0.976** | **0.939** | **0.623** | CLAM | 0.921 | 0.938 | 0.170 |

*Table 5: Embedding-level AUC, accuracy with 0.5 threshold, and specificity at 100% sensitivity ($Spec_{100}$) metrics for best MIL aggregation method for two OOD test sets reported for Scenarios A-D. Attn. = Attention-based aggregation; Gated = Gated attention; ACC = Accuracy.*




a)
b)
c)
d)
e)
f)
g)
h)
500µm
1mm
1
0
1
0




a)
0.9025
0.6730
0.6544
0.5992
0.5611
0.5586
0.5015
0.4939
0.4805
0.4692
b)
0.9551
0.5249
0.4951
0.4439
0.3976
0.3904
0.3894
0.3631
0.3494
0.3430
c)
1.0000
1.0000
0.9998
0.9990
0.9986
0.9986
0.9965
0.9959
0.9935
0.9929