scientific reports

OPEN

# Out-of-distribution generalization for segmentation of lymph node metastasis in breast cancer

Yiannis Varnava[1]✉, Kiran Jakate[8], Richard Garnett[1], Dimitrios Androutsos[1], Pascal N. Tyrrell[2,3,4] & April Khademi[1,2,4,5,6,7]

**Pathology provides the definitive diagnosis, and Artificial Intelligence (AI) tools are poised to improve accuracy, inter-rater agreement, and turn-around time (TAT) of pathologists, leading to improved quality of care. A high value clinical application is the grading of Lymph Node Metastasis (LNM) which is used for breast cancer staging and guides treatment decisions. A challenge of implementing AI tools widely for LNM classification is domain shift, where Out-of-Distribution (OOD) data has a different distribution than the In-Distribution (ID) data used to train the model, resulting in a drop in performance in OOD data. This work proposes a novel clustering and sampling method to automatically curate training datasets in an unsupervised manner with the aim of improving model generalization abilities. To evaluate the generalization performance of the proposed models, we applied a novel use of the Two One-sided Tests (TOST) method. This method examines whether the performance on ID and OOD data is equivalent, serving as a proxy for generalization. We provide the first evidence for computing equivalence margins that are data-dependent, which reduces subjectivity. The proposed framework shows the ensembled models constructed from models that generalized across both tumor and normal patches enhanced performance, achieving an F1 score of 0.81 for LNM classification on unseen ID and OOD samples. Interactive viewing of slide-level segmentations can be accessed on PathcoreFlow™ through https://web.pathcore.com/folder/18555?s=QTJVHJuhrfe5. Segmentation models are available at https://github.com/IAMLAB-Ryerson/OOD-Generalization-LNM.**

**Keywords** Histopathology, Lymph node, Breast cancer, Deep learning, Segmentation, Generalization

Over three hundred thousand new cases of breast cancer are expected to be diagnosed in 2024 among women in the United States, making it the most commonly diagnosed cancer within this demographic[1]. If tumor cells metastasize, the prognosis worsens, which increases the complexity of treatment and reduces the likelihood of survival. Pathologists evaluate lymph nodes to identify metastatic breast cancer, which can be labor-intensive and subjective. Computational pathology tools enabled by Artificial Intelligence (AI) can be used to overcome these challenges by offering objective and efficient metrics of disease to enhance quality of care[2]. AI adoption is becoming increasingly important as the increase in caseloads contribute to the mounting pressure on pathologists who are already grappling with heavy workloads[3]. The average pathologist workload has increased by 41.73% in the United States and 7.06% in Canada between 2007 and 2017[3], which may lead to pathologist burnout[4]. Therefore, AI can alleviate some of the workload burden on pathologists and reduce the risk of pathologist fatigue.

The "Cancer Metastases in Lymph Nodes Challenge" (CAMELYON) Lymph Node Metastasis (LNM) competitions were launched to compare automated algorithms for segmentation and detection of breast cancer tumor cells in lymph nodes[5–7]. The first challenge, CAMELYON16, focused on Whole-Slide Image (WSI) classification and metastasis detection while the second challenge, CAMELYON17, focused on the pathological N-stage (pN-stage) classification. These competitions have lead to the rapid advancement of

[1]Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada. [2]Department of Medical Imaging, University of Toronto, Toronto, ON, Canada. [3]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada. [4]Institute of Medical Science, University of Toronto, Toronto, ON, Canada. [5]Keenan Research Centre for Biomedical Science, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada. [6]Institute for Biomedical Engineering, Science Tech (iBEST), A Partnership Between St. Michael's Hospital and Toronto Metropolitan University, Toronto, ON, Canada. [7]Vector Institute for Artificial Intelligence, Toronto, ON, Canada. [8]Department of Pathology, Unity Health Toronto, Toronto, ON, Canada. ✉email: yvarnava@torontomu.ca

computational pathology tools for LNM segmentation and classification with deep learning tools presenting as top performers[8,9]. Despite the progress, several challenges persist, including poor generalization of deep learning models when applied to data that is outside the distribution of the training set. Domain shift is common in digital pathology imaging datasets and can be attributed to differences in scanner vendors, dataset compositions, staining variation, patient populations and more. This presents a fundamental barrier to wide-scale deployment and adoption, since data from new laboratories or sites can present as Out-of-Distribution (OOD) data and have lower performance[10]. Existing approaches for digital pathology aim to improve domain generalization with transfer learning, data balancing, hard-negative mining, image translation[11], and generative adversarial networks[12].

In this work, we propose an unsupervised method to curate a training dataset from a large number of patches to create patch-level models that generalize for LNM segmentation. The goal is to use more representative and balanced training datasets for patch-level segmentation to improve generalization. A novel clustering methodology based on pre-trained feature extractors from a histopathology dataset and ImageNet are utilized to balance datasets, and a sampling strategy is implemented to vary the number of partial tumor patches in the dataset. Top submissions to CAMELYON17 randomly sample a balanced amount of tumor and normal patches[13], or apply hard-negative mining[14,15] while others have grouped partial tumor patches into the tumor class[15,16]. Instead, we propose an automatic method of selecting the training dataset and compare the performance between the different models on both In-Distribution (ID) and OOD data to analyze generalization at the patch-level.

To compare and examine the performance of models across different data distributions, some studies assume a linear relationship between ID and OOD performance[17–19], while others report mixed findings[20,21]. Taori et al. examined the robustness of ImageNet models to distribution shifts caused by natural changes within datasets and found a significant performance gap between ID and OOD data[17]. The authors defined a measure of "Effective Robustness" as a proxy for generalization based on the assumption of a linear relationship between ID and OOD performance. Research from Andreassen et al.[18] investigates the evolution of OOD robustness during the fine-tuning of deep learning models, discovering that early stopping can help maintain a balance between ID and OOD performance. Wenzel et al. explored the impact of fine-tuning on OOD generalization and found the relationship between ID and OOD performance depends heavily on the specific datasets used[20]. Teney et al.[21] suggested there is a trade-off between optimizing for ID performance and preserving OOD generalization, noting that ID and OOD performance may sometimes be inversely correlated, depending on the dataset.

Here, we propose a novel application of the Two One-sided Tests (TOST)[22], for examining model generalization through investigating the statistical equivalence of performance in ID and OOD datasets for LNM segmentation. Our proposed method for determining equivalence boundaries within the TOST framework is grounded in a data-driven approach, using statistics such as the mean and standard error directly from the data, thus reducing subjectivity. By relying on the data itself to guide the boundary selection, our methodology enhances robustness and adaptability across different clinical and modeling scenarios. This work provides the first evidence supporting the computation of data-dependent equivalence margins, as opposed to using predetermined or arbitrary thresholds. This first application of TOST for generalization analysis could have wide applicability to many digital pathology and computer vision tasks. The CAMELYON datasets, in addition to a private clinical dataset of 58 WSI are used in this study to test this framework. The held-out ID and OOD testing datasets comprise 237 WSI and 1,393,890 patches. As we show, the ensemble models that generalize both on tumor and normal patches separately have the top LNM classification performance.

## Materials and methods

This work focuses on developing patch-level tumor segmentation models that are robust and generalize to OOD data. The patch-level segmentation pipelines for training and testing are shown in Fig. 1. As opposed to comparing different model architectures, we compare the same architecture (DeepLabV3[23]) on various datasets that are curated using a novel sampling and clustering technique. To determine which models generalize to OOD data, a new application of TOST analysis is applied to examine the performance across held-out datasets. Patch-level masks are then recomposed into whole-slide prediction masks, and evaluated on the testing set for LNM classification. It is hypothesized that unsupervised training data selection using cluster analysis can improve generalization.

### Datasets

There are 507 WSIs used in this work, comprised of two open source datasets and a clinical dataset. Four LNM stages are included: Macro-Metastasis (macro), Micro-Metastasis (micro), Isolated Tumor Cells (ITC), and negative. See Table 1 for sample sizes and Table 2 for dataset properties. The official CAMELYON16 training slides are used to train all patch-level segmentation models, which differ from one another based on the training dataset selected by the proposed sampling and clustering techniques. To examine generalization performance of the patch-level segmentation models, the testing dataset contains both (held-out) ID and OOD data. The ID data includes all WSI scanned at the same centers as the CAMELYON16 training slides. The OOD test data includes a distinct center from CAMELYON17 and a clinical dataset from St. Michael's Hospital (SMH). The annotations from CAMELYON16 and CAMELYON17 are provided by the competition, and for the clinical SMH dataset, a breast pathologist (K.J) labeled the annotated tumors using PathcoreFlow™[24]. A breakdown of ID and OOD WSI, grouped by LNM is shown in Supplementary Table S1. All WSIs in this study lack any patient-identifying information.

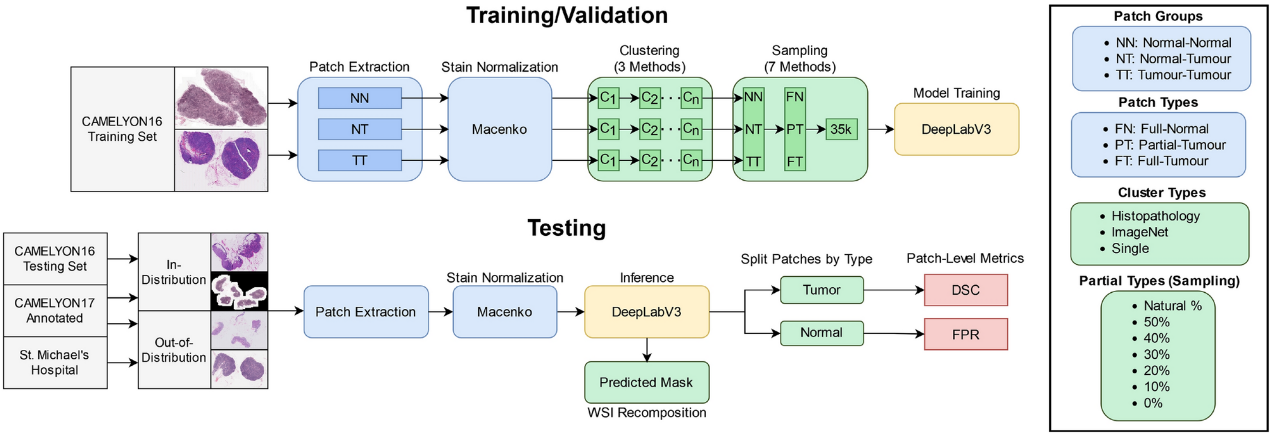

**Fig. 1**. Training and inference pipelines for patch-level segmentation models. During training, features are separately extracted from images belonging to each patch group before being clustered. We evenly sample from each cluster ($C_1, C_2, ...C_n$) with the goal of obtaining the desired percentage of partial tumor patches in the dataset.

| LNM | CAMELYON16 train | CAMELYON16 test | CAMELYON17 | SMH | Total |
|---|---|---|---|---|---|
| Macro-metastasis | 58 | 22 | 17 | 4 | 101 |
| Micro-metastasis | 53 | 27 | 17 | 18 | 115 |
| Isolated tumor cells | 0 | 0 | 16 | 6 | 22 |
| Negative | 159 | 80 | 0 | 30 | 269 |
| Total | 270 | 129 | 50 | 58 | 507 |

**Table 1**. Summary of all WSI used for training and testing from each dataset, grouped by LNM stage.

| Dataset | Center | Scanner | Resolution (µm/px) | Magnification | ID/OOD | # WSI (Train/Test) |
|---|---|---|---|---|---|---|
| CAMELYON16 | RUMC* | 3D Histech P250 | 0.24 | 40× | ID | 171/74 |
| CAMELYON16 | UMCU* | Hamamatsu NanoZoomer-XR C12000-01 | 0.23 | 40× | ID | 99/55 |
| CAMELYON17 | CWZ** | 3D Histech P250 | 0.24 | 40× | ID | 0/10 |
| CAMELYON17 | RST** | 3D Histech P250 | 0.24 | 40× | ID | 0/10 |
| CAMELYON17 | UMCU* | Hamamatsu NanoZoomer-XR C12000-01 | 0.23 | 40× | ID | 0/10 |
| CAMELYON17 | RUMC* | 3D Histech P250 | 0.24 | 40× | ID | 0/10 |
| CAMELYON17 | LPON | Philips Intellisite Ultra Fast Scanner | 0.25 | 40× | OOD | 0/10 |
| SMH | SMH | Aperio ScanScope | 0.50 | 20× | OOD | 0/58 |

**Table 2**. Overview of the datasets used in the study for training and testing. The table also indicates whether the data is ID or OOD. Centers marked with * exist in both CAMELYON16 and CAMELYON17, and centers marked with ** sent WSIs to RUMC for scanning. *RUMC* Radboud University Medical Center, Nijmegen; *UMCU* University Medical Center Utrecht; *CWZ* Canisius-Wilhelmina Hospital, Nijmegen; *RST* Rijnstate Hospital, Arnhem; *LPON* Laboratorium Pathologie Oost-Nederland, Hengelo; *SMH* St. Michael's Hospital, Toronto, Canada.

## Image pre-processing

For training, OpenSlide[25] is used to read and extract patches of $256 \times 256$ pixels from WSIs at 20x magnification. Macenko[26] normalization from the Tissue Image Analytics toolbox[27] is applied to normalize color distributions of slides acquired from different medical centers, necessary for accurate analysis in digital pathology[28,29]. To make patch extraction more efficient, only WSI regions with a considerable amount of tissue are extracted for further analysis. Similar to[13,30,31], each patch is converted to the HSV color space and Otsu's thresholding[32] is used to detect foreground. For normal patches, Otsu's thresholding method is used to remove patches with large amounts of background (at least 10% of the patch must have nonzero intensity). Images with 10% tumor (or more) within a patch are labelled as tumor and retained for the tumor class. Segmentation models are trained from Macenko normalized patches augmented with horizontal and vertical flips with probability $p = 0.5$, as well as color jitter ($p = 0.1$) on the contrast and saturation.

## Patch groups

Patches are split into three "patch groups": normal–normal (NN), normal–tumor (NT), and tumor–tumor (TT). NN refers to patches from slides labeled as normal (contain no tumor). NT refers to normal patches from slides with tumors, and TT refers to patches with tumorous cells. We adopt a similar method as Sikaroudi et al.[33] to create the pool of extracted patches, which involves extracting the same amount of tumor and normal patches from a respective slide. However, we also utilize normal patches from the normal slides. We begin by randomly extracting as many non-overlapping tumor patches (TT) as possible since tumor patches are the limiting factor. For NT patches, the same amount of patches are extracted from those tumorous slides. Finally, for NN patches, we extract 500 normal patches from each slide, which is roughly the same amount of NT patches extracted. Overall, this creates approximately 230, 000 patches from 270 CAMELYON16 training slides to be clustered and sampled. We construct our training datasets by sampling from each cluster for each patch group.

## Clustering

We investigate unsupervised image clustering methods for optimizing training datasets and observe the effects of different training set compositions on ID and OOD data. Unsupervised clustering can categorize training patches into subgroups based on image characteristics. We independently cluster each of the patch groups (NN, NT, TT) to obtain clusters of patches for each, which are then sampled by the patch type. In this work, we experiment with two different feature extraction (clustering) methods and one baseline approach. The first two clustering methods are based on extracted features from ResNet-18[34] models pre-trained on (1) natural images from ImageNet[35], and (2) histopathology data[36]. Specifically, the "pytorchnative_tenpercent_resnet18.ckpt" weights are used for the histopathology model[36]. The ResNet-18-ImageNet model was trained using supervised learning while the ResNet-18-Histopathology model was trained using self-supervised learning. The image patches from the training dataset are supplied to each of the models, and the features are clustered using a Gaussian Mixture Model. The optimal number of clusters are determined by the Elbow method[37–39] using the Calinksi-Harabasz (CH) Index[40] and the Davies-Bouldin (DB) Index[41]. The CH index measures the ratio of the between-cluster variance to the within-cluster variance, while DB seeks to minimize the variance within clusters while simultaneously maximizing the separation distance between clusters. To find the optimal number of clusters, a sweep of the number of clusters was conducted, ranging from 2 to 10, with the optimal number selected using the Elbow method[37] which is a standard approach in cluster analysis[38,39]. The elbow method in cluster analysis involves plotting the CH index and/or DB index against a number of clusters and selecting the point where the rate of decrease (or increase, depending on the metric) sharply changes, forming an "elbow", which suggests the optimal number of clusters. Both the CH and DB metrics were used in this work. This approach looks for similar features in subgroups, resulting in more homogeneous representations for the selected patches. We hypothesize that balancing the dataset using novel feature clustering strategies may help with generalization. The baseline comparison is treating the entire population of patches as a single cluster (Single).

## Sampling

After applying clustering to each NN, NT and TT patch group separately, we stratify sampling by patch type across each cluster for each patch group to construct a final dataset of around 35, 000 patches. Patch type is defined based on tumor content: Full-Tumor (FT)—100% tumor, Full-Normal (FN)—100% normal, and Partial-Tumor (PT)—a nonzero percentage of tumor. The result is a dataset with approximately balanced NN, NT, and TT patches, stratified by FT, FN, and PT, across the different clusters. The amount of PT patches varies in the final datasets, and the "Partial Type" is the percentage of PT patches: 0%, 10%, 20%, 30%, 40%, 50%, and Natural%. The Natural% partial type refers to randomly sampled patches from each patch group, with the majority being FN (most commonly occurring tissue/patch), and the rest being a combination of FT and PT patches. Supplementary Table S2 contains the exact percentage of patch types in each sampled dataset. In total, clustering and sampling creates $3 \times 7 = 21$ training datasets (3 clustering types and 7 partial types).

## Patch-level segmentation

Patch-level segmentation models are developed using a pre-trained (ImageNet) PyTorch implementation of the DeepLabV3 architecture[23] with a MobileNetV3-Large backbone[42]. We apply transfer learning to adapt the model to our dataset. The segmentation is binary and pixel-level, where the tumor and normal classes are the positive and negative classes, respectively. The Focal Tversky Loss[43], based on the Tversky Index[44], is used to focus on difficult examples ($\gamma$) while also minimizing false positive predictions ($\alpha$ and $\beta$) with $\alpha = 0.3, \beta = 0.7$ , and $\gamma = \frac{1}{2}$. We utilize this loss function for its ability to emphasize harder examples during training (Focal), while also providing the flexibility to adjust the weighting of false positives and false negatives (Tversky).

## Equivalence testing for generalization

We present a novel method for analyzing the generalization capabilities of segmentation models for digital pathology which considers generalizing models to have similar patch-level segmentation performance on ID and OOD data. To determine whether performance is equivalent in ID and OOD data (i.e., generalizes), we utilize TOST, a form of statistical equivalence test used in clinical trials to determine treatment effect[22]. TOST is leveraged as opposed to a t-test since we are testing for performance equivalence, and not if the difference between performance in ID and OOD is statistically significant. This is a new application of TOST for generalization analysis.

Let $\mu_1$ and $\mu_2$ be the mean performance metrics on held-out ID and OOD data, respectively, computed for each of the 21 models. TOST analysis is used to determine whether the performance on ID and OOD data are statistically equivalent if the observed difference, $\mu_1 - \mu_2$, falls within a specified equivalence interval $(-\Delta, \Delta$

). The lower (L) and upper (U) null hypotheses ($H_{0,L}$ and $H_{0,U}$) and alternative hypotheses ($H_{1,L}$ and $H_{1,U}$) are defined as follows:

$$H_{0,L} : \mu_1 - \mu_2 \leq -\Delta, \quad H_{0,U} : \mu_1 - \mu_2 \geq \Delta,$$
$$H_{1,L} : \mu_1 - \mu_2 > -\Delta, \quad H_{1,U} : \mu_1 - \mu_2 < \Delta.$$

To determine if the performance is equivalent in ID and OOD data (i.e., the model generalizes), the entire 90% confidence interval (CI) for the difference $\mu_1 - \mu_2$ must lie within the equivalence interval. This also means that both null hypotheses need to be rejected. Our study applies this method on each model $i$, where $i$ ranges from 1 to $n$ and $n = 21$. The mean performance produced by model $i$ on ID data is denoted as $\mu_{1,i}$, and the mean performance on OOD data is denoted as $\mu_{2,i}$. Therefore, the difference in means for each model, $D_i$, is computed as: $D_i = \mu_{1,i} - \mu_{2,i}$.

To determine the equivalence interval and bounds, the standard error is used to calculate the 95% confidence interval. The mean ($\mu_D$), standard deviation ($\sigma_D$), and standard error ($\mathrm{SE}_D$) of the differences are found by:

$$\mu_D = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad \sigma_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \mu_D)^2}, \quad \mathrm{SE}_D = \frac{\sigma_D}{\sqrt{n}},$$

and the equivalence interval ($\pm\Delta$) is computed as:

$$\Delta = |\mu_D| + |1.96 \times \mathrm{SE}_D|\,,$$

where we add a tolerance around the mean difference based on the standard error.

The upper and lower bounds are designed to be symmetrical around a mean difference of 0, supporting the ideal scenario where the performance on ID and OOD data is the same. Equivalence bounds are calculated separately for the tumor patches and the normal patches due to different metrics being used (DSC versus FPR), as well as for each LNM stage, due to performance variations (macro > micro > ITC). The TOSTtwo.raw function from the TOSTER R package is used to perform the equivalence tests with an alpha of 0.05[45,46]. We have not found any studies using TOST for generalization analysis, and we propose this as a novel method for this task.

## Classification

The LNM classification pipeline is shown in Supplementary Fig. S1. For each model, patch-level segmentation predictions are used to recompose the corresponding WSI. Segmentation performance at the WSI-level is examined by downsampling the images and predictions by 16x. For WSI-level classification of LNM category, a clinically-informed rule-based method is employed that mimics the pathologists scoring of lymph node tissue. The method is part of the Tumor, Node, Metastasis (TNM) staging system[47] which uses the diameter of the largest tumor in the lymph node tissue, and assigns an LNM category as in Table 3. To classify WSI, each object in the recomposed predictions is identified using connected components and the major axis length of a fitted ellipse is computed to measure the diameter. The largest diameter in the WSI is used to predict the LNM category. No post-processing is employed. Performance is also examined in terms of "positive" and "negative" classes to mimic clinical decision-making[47–49].

In addition to individual model predictions, the result of ensembling several models is also investigated. The ensemble configurations include: Generalizing (models found to generalize using TOST), Non-Generalizing (models not in the equivalence interval for TOST), Generalizing (T&N) for models that generalize on both tumor and normal, Generalizing (T) for tumor, Generalizing (N) for normal, Natural%, 50%, 40%, 30%, 20%, 10%, 0%, Histopathology, ImageNet, and Single. To analyze classification performance of the 21 individual models and the 15 ensembled predictions, results were generated using a majority vote from the WSI-level predictions over all folds for a particular model.

## Validation metrics

Segmentation performance is quantified using the DSC on the tumor patches and WSI-level:

| LNM | Description | Class |
|---|---|---|
| Macro-metastasis | $x > 2$ mm | Positive |
| Micro-metastasis | 0.2 mm $< x \leq 2$ mm | |
| Isolated tumor cells | $x \leq 0.2$ mm or $\leq 200$ cells | Negative |
| Negative | No tumor found | |

**Table 3.** LNM categories, where *x* represents the largest dimension of the tumor.

$$\mathrm{DSC} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

where TP, FP, and FN represent the true positive, false positive, and false negative pixels, respectively. It is not possible to use the DSC metric on normal patches as there are no true positives. We believe performance on normal patches is largely unreported in the literature. To quantify performance on normal patches (patches with no tumor pixels) we use only the False Positive Rate (FPR)

$$\mathrm{FPR} = 1 - \mathrm{Specificity} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{2}$$

For WSI-level classification performance, the LNM category is assigned based on the largest diameter and performance is measured using the True Positive Rate (TPR/Sensitivity/Recall),

$$\mathrm{TPR} = \mathrm{Sensitivity} = \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

which highlights the number of correctly classified WSI, as well as 1-Specificity (Eq. 2) and Precision

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4}$$

which is the ratio of correctly predicted positive cases to all predicted positive cases. We also use the F1 score which is a tradeoff between precision and recall (sensitivity). All metrics in this study are reported as the mean value over 5 folds unless otherwise specified.

## Results

For training the segmentation models, the Adam optimizer with an initial learning rate of $1e^{-4}$ and an exponential decay are used. Each model is set to train with a batch size of 8 for 50 epochs with early stopping. All models are trained on a single NVIDIA V100 GPU (32GB) through the Digital Research Alliance of Canada. We apply 5-fold cross validation for each of the 21 model combinations to train a total of 105 models. We hold out unseen ID and OOD WSI to test our models and aggregate performance metrics across the folds to obtain results. Inference is performed on a NVIDIA RTX 3090 (24GB) GPU for all models.

### Clustering

The optimal number of clusters for each combination of feature extractor and patch group are as follows for ImageNet: NN: 6, NT: 7, TT: 5. For histopathology, the optimal number of clusters are: NN: 6, NT: 7, TT: 3. The clusters are used to gather patches with similar characteristics for further sampling. Supplementary Figs. S2–S4 illustrate clusters of patches from the three feature extraction methods-Histopathology, ImageNet, and Single, respectively, for each patch group, highlighting the unique patterns in the unsupervised training data selection. Figures 2 and 3 show the clusters in 2D scatter-plots along with their respective CH and DB indices for all patch groups from the Histopathology and ImageNet methods, respectively. The chosen number of clusters is indicated as the red point. The 2D visualizations can be used to associate which types of patches are included in each cluster, while the CH and DB indices along with the elbow method are used to choose the optimal number of clusters. Both methods (Histopathology and ImageNet) found the same number of clusters for each patch group except in the TT patch group. In each respective case, the number of clusters that resulted in less cluster overlap was chosen based on the CH and DB indices.

### Patch-level segmentation performance

Qualitative results for patch-level segmentation are shown in Fig. 4, demonstrating good performance across various patch compositions. Supplementary Figs. S5 and S6 show examples where the models were performing poorly. For quantiative performance analysis, the mean ID and OOD performance for each model on tumor and normal patches is shown in Table 4. The best performing model in tumor patches is Histopathology-40%, with mean DSC $= 0.75$). Histopathology-Natural% and Single-Natural% are the top performers for normal patches with the lowest FPRs. Figure 5 shows the mean performance over tumor and normal patches, as a function of PT and clustering type. Models trained on datasets with higher PT generally perform better on tumor whereas models trained on datasets with lower PT generally perform better on normal. Natural sampling has the lowest FPR in normal patches. Histopathology clustering has the best DSC in tumor patches and Single has the lowest FPR in normal patches. Interestingly, ImageNet clustering has the lowest DSC and highest FPR. The mean ID and mean OOD performance on each LNM is represented as a heatmap in Supplementary Fig. S7. The scatter-plot shown in Fig. 6 illustrates the trade-off between DSC and FPR at the patch-level for each model.

### Generalization analysis for patch-level segmentation

For patch-level segmentation, TOST results are summarized for tumor (DSC) and normal (FPR) patches in Table 4, highlighting the computed equivalence bounds, difference in means ($D$), 90% CI, and p-value used to determine whether a given model has OOD performance within the equivalence interval of the ID data ($p < 0.05$). TOST results are visualized in Fig. 7 for tumor/normal and in Supplementary Fig. S8 for each

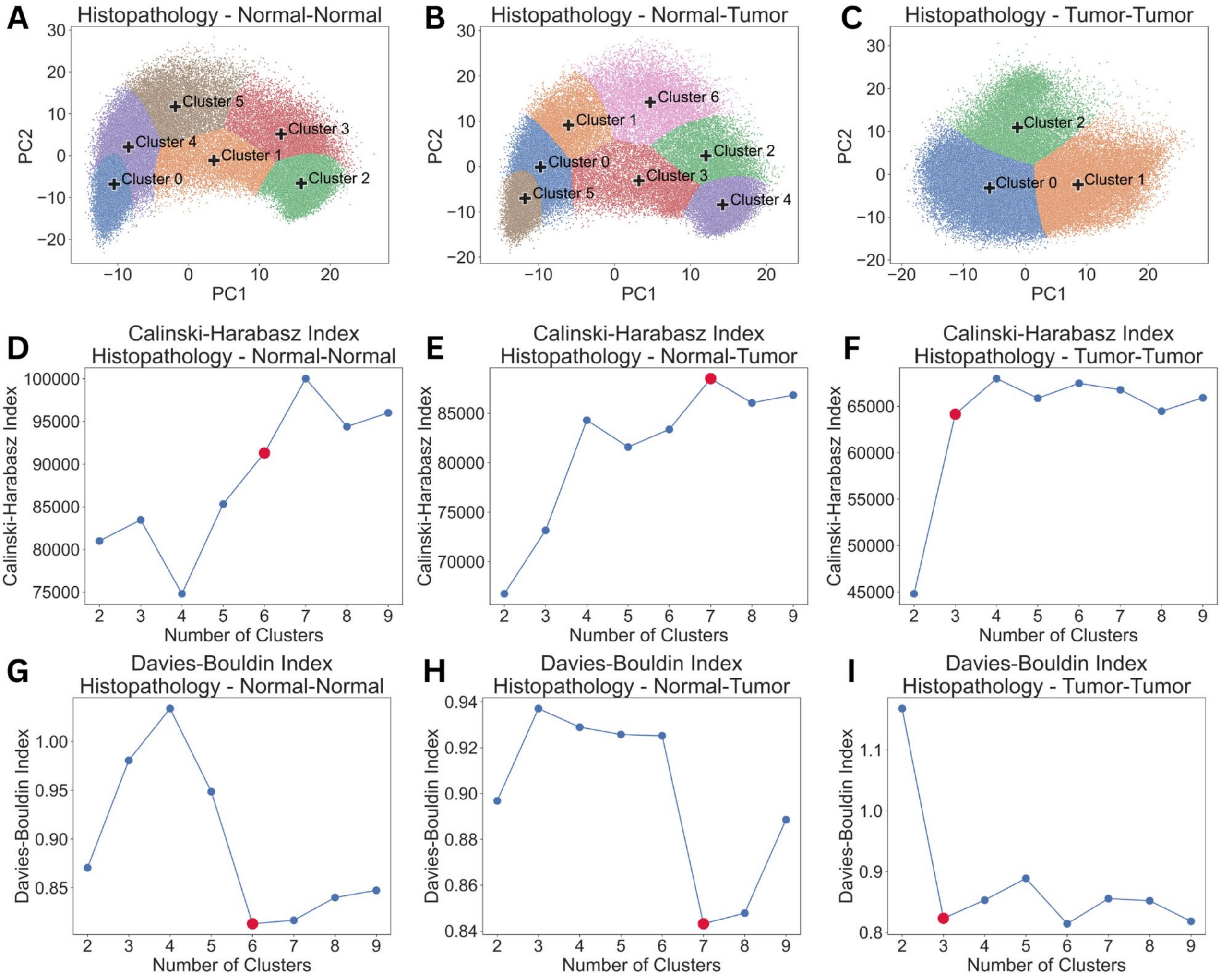

**Fig. 2**. 2D features and clusters (**A**–**C**), Calinski–Harabasz Index (**D**–**F**)), and Davies–Bouldin Index (**G**–**I**) results for multiple clusters for each patch group using the Histopathology feature extractor. The optimal number of clusters is shown as a red point.

LNM. Models that generalize in tumor patches are Histopathology-0%, Histopathology-40%, ImageNet-0%, ImageNet-50%, and Single-Natural%. Histopathology clustering has the most number of models generalizing over LNM categories. The generalizing models on normal patches include Histopathology-0%, Histopathology-Natural%, ImageNet-Natural%, Single-30%, and Single-Natural%. That is, all three Natural% models generalize on normal patches.

## WSI-level lymph node metastasis classification

Patch-level model predictions on all held-out ID and OOD data are used to recompose each WSI. The Generalizing ensemble contains all 8 generalizing models (on either tumor or normal patches), and for a fair comparison, the Non-Generalizing ensemble contains a randomly selected 8 out of 13 models that did not generalize. Supplementary Table S3 details the models used in each ensemble. WSI-level segmentation performance on macro, micro, and ITC (DSC, tumor) and on negative (FPR, normal) slides for all models in Supplementary Figure S9 and by LNM in Supplementary Fig. S10 show that compared to individual models, the ensemble models have higher performance. The Generalizing ensemble has the highest DSC and Generalizing (T&N) has the lowest FPR.

Using the largest diameter of the detected objects in the WSI, the LNM category is predicted using positive and negative slide labels. Tables 5 and 6 show the classification performance of all individual and ensemble models on LNM classification across positive and negative WSI. Single-Natural% is a top performer in the individual models with similar performance in Histopathology-Natural%. In the ensembles, Generalizing (T&N) is the best performing across all metrics and both individual and ensemble models, followed by Generalizing and Generalizing (N). Most notably, the F1 score is substantially higher than the best Single model (F1 = 0.805 versus F1 = 0.680) which shows the power of using generalizing models for WSI analysis of LNM. Multi-class LNM classification results is shown in Supplementary Table S4 for individual models and Supplementary Table S5 for

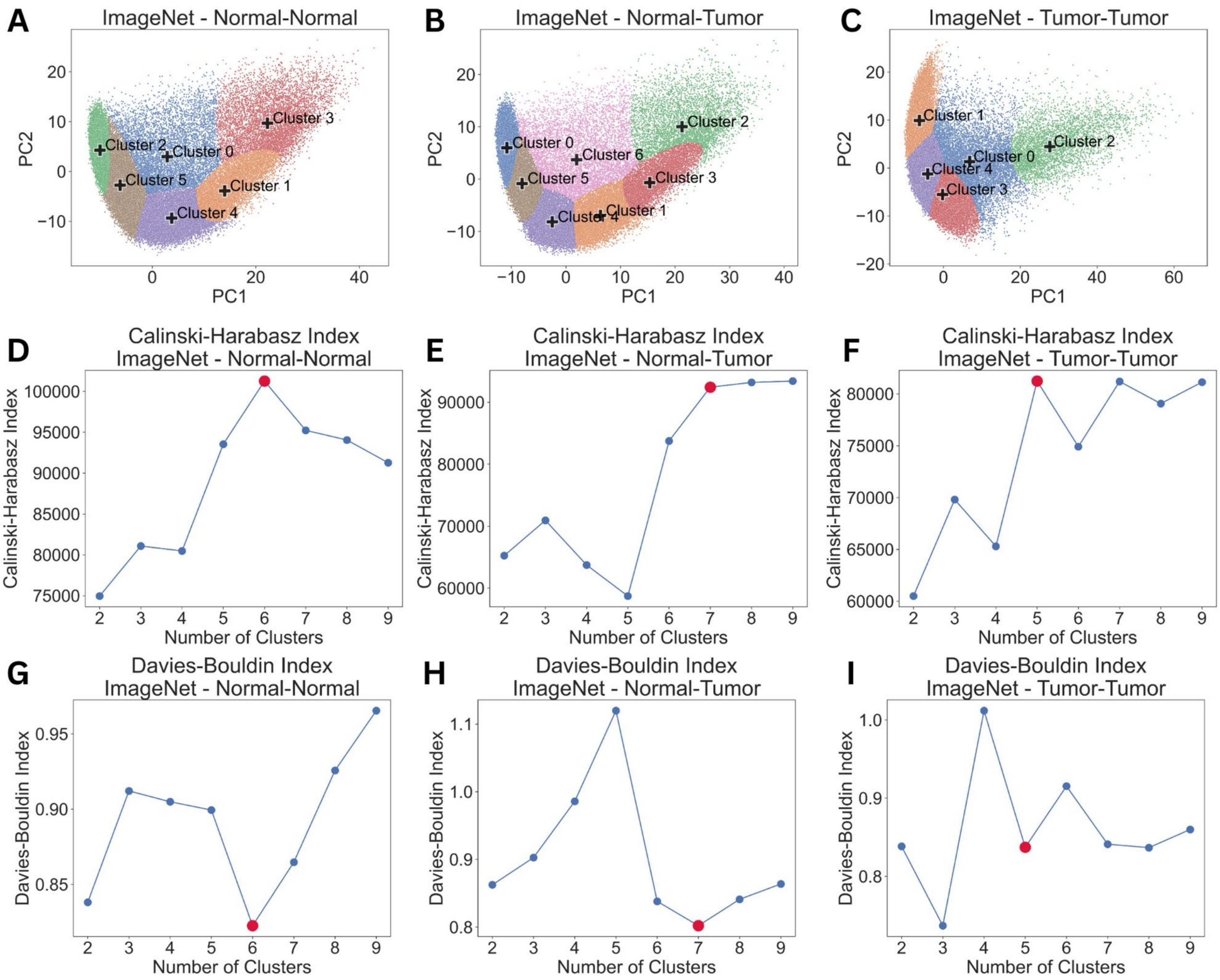

**Fig. 3**. 2D features and clusters (**A**–**C**), Calinski–Harabasz Index (**D**–**F**)), and Davies–Bouldin Index (**G**–**I**) results for multiple clusters for each patch group using the ImageNet feature extractor. The optimal number of clusters is shown as a red point.

ensemble models. Additionally, Supplementary Tables S6 (individual) and S7 (ensemble) contain the results, split by individual LNM labels, which exhibits similar trends.

To analyze the results further, the confusion matrices of the top models are shown in Fig. 8 for the positive/negative classes. The benefits of using the generalizing models becomes more apparent when looking at individual class performance. The Single-Natural% and Non-Generalizing ensemble exhibit high positive classification accuracy (95.0% and 91.8%), but low negative classification accuracy (45.8% and 49.5%). Conversely, both the Generalizing and Generalizing (T&N) ensembles both produce high sensitivity on the positive classes (90.9% and 84.6%), while also having better performance on the negative classes (62.2% and 76.7%) compared to Single-Natural% and Non-Generalizing. When considering the confusion matrices for multi-class LNM classification in Supplementary Fig. S11, Generalizing and Generalizing (T&N) have slightly lower performance in the macro and micro classes, with much higher performance across ITC and negative.

## Discussion

This work presents a comprehensive framework for testing the generalization capabilities of patch-level LNM segmentation in WSI. Previous works primarily focused on the performance of WSI-level classification only, but for increased explainability of AI results, and for optimizing models, we rigorously examine patch-level segmentation performance in ID and OOD data. We propose cluster-based sampling strategies to balance datasets in an unsupervised manner, and a novel application of the TOST statistical testing, along with data-driven margins to determine which models are generalizing. In clinical contexts where performance criteria are predefined, such as in clinical trials with established benchmarks, the proposed method can easily accommodate fixed margins. However, in areas lacking such standards, particularly for generalization capabilities, our statistical approach will allow for more reliable determination of equivalence boundaries based on observed data. This overcomes some of the inherent limitations of traditional TOST when applied to complex, real-world scenarios. TOST is expected to be broadly applicable to generalization tasks in machine learning and other areas due to

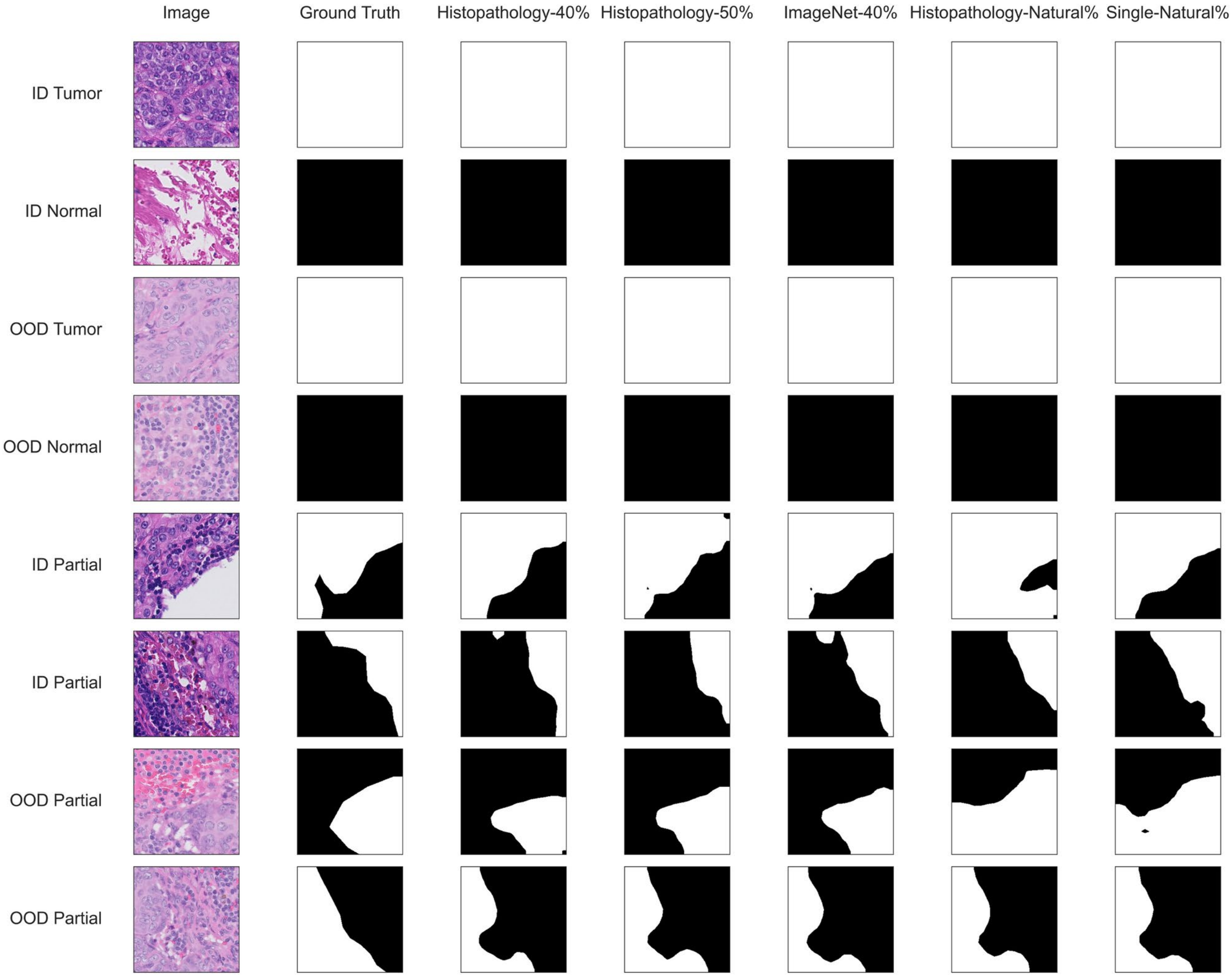

**Fig. 4**. Visual results displaying the original image patch, ground truth, and prediction mask from the top generalizing model for each LNM: Histopathology-40% (tumor and macro), Histopathology-50% (micro), ImageNet-40% (ITC), Histopathology-Natural% and Single-Natural% (normal). Single-Natural% is shown as the baseline model as well. Full-tumor, partial-tumor, and full-normal patches are shown from one ID WSI and one OOD WSI.

its adaptability and automated boundary derivation. In the analysis, common segmentation metrics (DSC for tumor patches, FPR for normal patches, sensitivity, 1-specificity, precision, and F1 score) were used to provide a comprehensive evaluation.

At the patch-level, 21 individual models were developed using ID training data selected through cluster-based sampling techniques that were then applied to held-out ID and OOD data. ID data is from the CAMELYON challenges, and OOD data is from a private clinical dataset. In total, there are 1,393,890 patches in the held out dataset, where 104,842 contain tumor and 1,289,048 are completely normal. Due to false positives being a significant challenge when reconstructing WSI from patch-level predictions, we opted to not only report on tumor segmentation performance but also normal patch performance. TOST revealed several models with equivalent ID and OOD performance which implies that these models are generalizing. Ensembling together the two models that generalized in both normal and tumor patches led to a substantial gain in WSI classification performance compared to baseline (Single-Natural%), which would lead to improved patient management.

We find importance in the cluster type for generalization. Histopathology is the only cluster type in the study that has at least one model that generalizes for each LNM which indicates that this feature extractor may be aiding in model robustness and generalization. There are both similarities and differences associated with the types of patches selected in each cluster for Histopathology (Supplementary Fig. S2) and ImageNet (Supplementary Fig. S3) pre-trained feature extractors, specifically relating to the diverse tissue types found in the data (Supplementary Fig. S12). By clustering similar patches like histiocytes, stroma, and artifacts into their own groups, the training dataset could end up with more samples from these types of patches, potentially enhancing the model's robustness. The tumor-tumor patch group contains 3 clusters for Histopathology and 5 for ImageNet. The Histopathology clusters vary by tumor cell size and density, while for ImageNet there is more randomness in the clusters, which is undesirable. We attribute this to be a potential cause for decreased DSC

| | | Patch-level segmentation | | Equivalence testing | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Tumor | | | Normal | | |
| | | Tumor | Normal | **− 0.079, 0.079** | | | **− 0.006, 0.006** | | |
| **Cluster type** | **Partial type** | **Mean DSC** | **Mean FPR** | ***D*** | **90% CI** | **p value** | ***D*** | **90% CI** | **p value** |
| Histopathology | 0 | 0.72 (0.43) | 0.004 (0.060) | 0.000 | − 0.028, 0.027 | < **0.001** | 0.003 | 0.002, 0.005 | **0.019** |
| Histopathology | 10 | 0.73 (0.40) | 0.006 (0.064) | 0.052 | 0.002, 0.103 | 0.169 | 0.005 | 0.003, 0.007 | 0.253 |
| Histopathology | 20 | 0.73 (0.39) | 0.007 (0.068) | 0.066 | 0.039, 0.092 | 0.175 | 0.006 | 0.004, 0.007 | 0.331 |
| Histopathology | 30 | 0.71 (0.39) | 0.007 (0.064) | 0.098 | 0.081, 0.115 | 0.963 | 0.006 | 0.003, 0.008 | 0.402 |
| Histopathology | 40 | **0.75 (0.37)** | 0.010 (0.076) | 0.017 | − 0.004, 0.038 | < **0.001** | 0.004 | 0.001, 0.008 | 0.219 |
| Histopathology | 50 | 0.70 (0.39) | 0.012 (0.084) | 0.103 | 0.041, 0.166 | 0.776 | 0.005 | 0.003, 0.007 | 0.294 |
| Histopathology | Natural | 0.68 (0.42) | **0.003 (0.047)** | 0.070 | 0.041, 0.098 | 0.279 | 0.003 | 0.001, 0.005 | **0.015** |
| ImageNet | 0 | 0.71 (0.43) | 0.006 (0.072) | 0.026 | − 0.010, 0.063 | **0.016** | 0.006 | 0.003, 0.008 | 0.459 |
| ImageNet | 10 | 0.71 (0.40) | 0.008 (0.077) | 0.089 | 0.046, 0.131 | 0.672 | 0.008 | 0.003, 0.014 | 0.804 |
| ImageNet | 20 | 0.71 (0.40) | 0.006 (0.062) | 0.082 | 0.054, 0.109 | 0.577 | 0.006 | 0.004, 0.007 | 0.393 |
| ImageNet | 30 | 0.71 (0.39) | 0.009 (0.074) | 0.089 | 0.007, 0.172 | 0.598 | 0.007 | 0.003, 0.011 | 0.72 |
| ImageNet | 40 | 0.67 (0.40) | 0.010 (0.078) | 0.171 | 0.099, 0.243 | 0.974 | 0.008 | 0.005, 0.011 | 0.856 |
| ImageNet | 50 | 0.73 (0.38) | 0.010 (0.075) | 0.041 | 0.012, 0.069 | **0.022** | 0.009 | 0.005, 0.013 | 0.917 |
| ImageNet | Natural | 0.71 (0.41) | 0.004 (0.053) | 0.043 | − 0.002, 0.088 | 0.082 | 0.004 | 0.003, 0.005 | **0.002** |
| Single | 0 | 0.70 (0.44) | 0.004 (0.060) | 0.041 | − 0.013, 0.095 | 0.110 | 0.003 | 0.000, 0.006 | 0.059 |
| Single | 10 | 0.71 (0.40) | 0.005 (0.056) | 0.056 | 0.009, 0.104 | 0.190 | 0.005 | 0.003, 0.006 | 0.081 |
| Single | 20 | 0.73 (0.39) | 0.006 (0.062) | 0.047 | 0.002, 0.091 | 0.099 | 0.004 | 0.002, 0.006 | 0.101 |
| Single | 30 | 0.73 (0.38) | 0.006 (0.057) | 0.059 | 0.032, 0.086 | 0.101 | 0.004 | 0.002, 0.006 | **0.047** |
| Single | 40 | 0.70 (0.39) | 0.007 (0.061) | 0.086 | 0.049, 0.124 | 0.648 | 0.004 | 0.001, 0.007 | 0.131 |
| Single | 50 | 0.73 (0.38) | 0.011 (0.074) | 0.058 | 0.018, 0.099 | 0.178 | 0.005 | 0.002, 0.008 | 0.304 |
| Single | Natural | 0.72 (0.40) | **0.003 (0.045)** | 0.029 | − 0.006, 0.064 | **0.017** | 0.002 | 0.001, 0.004 | < **0.001** |

**Table 4**. Mean DSC (tumor) and mean FPR (normal) across ID and OOD datasets for all models, reported as mean (SD). TOST results for difference in means, *D*, between ID and OOD datasets by fold, with 90% CIs and p-values. Top-performing models and significant values are in bold.

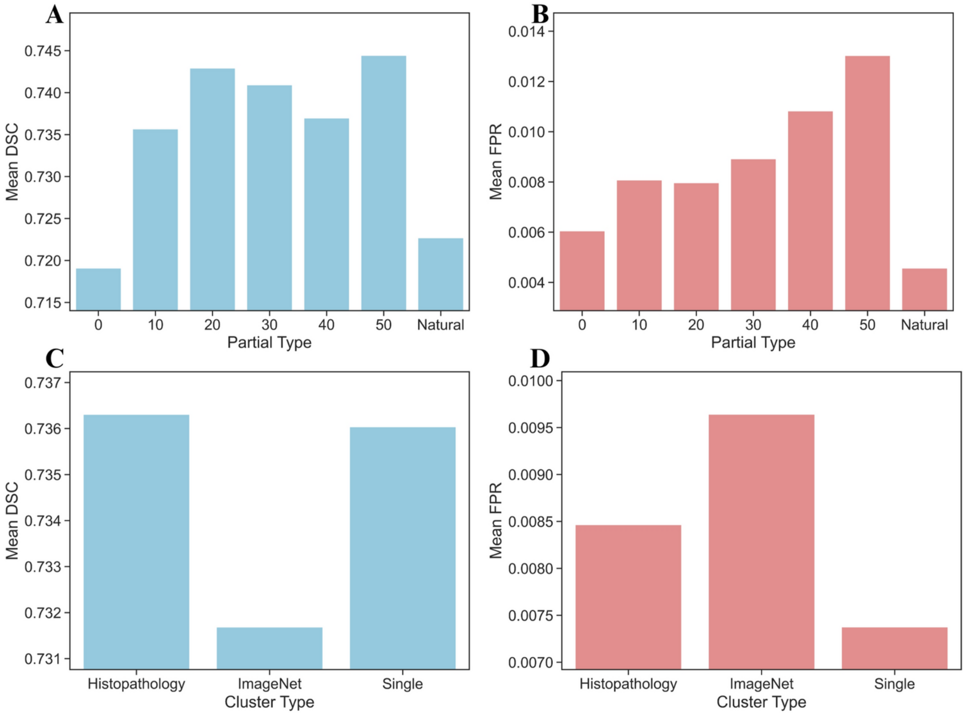

**Fig. 5**. Patch-level segmentation results showing Mean DSC for tumor patches and Mean FPR for normal patches across different groupings. (**A**) Tumor DSC by partial type. (**B**) Normal FPR by partial type. (**C**) Tumor DSC by cluster type. (**D**) Normal FPR by cluster type.

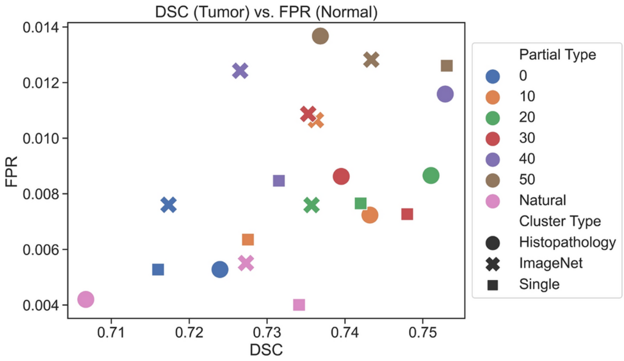

**Fig. 6**. Scatter-plot of the mean DSC (tumor) versus mean FPR (normal) across the entire testing set for all models.

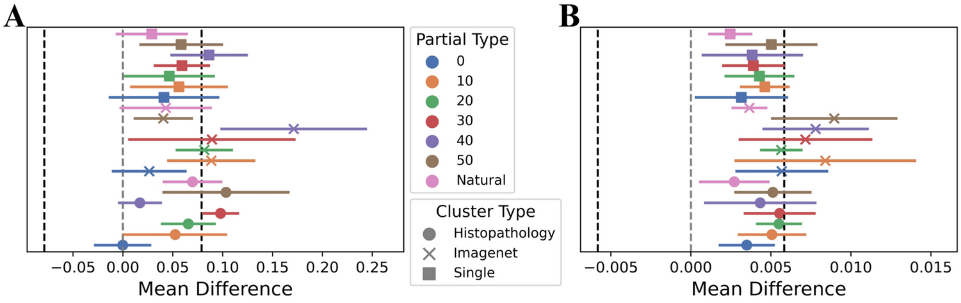

**Fig. 7**. Difference in means ($D$) and 90% CI for (**A**) tumor patches, and (**B**) normal patches. Models are considered equivalent if the CI is within the equivalence bounds ($p < 0.05$), indicated by the black vertical dashed lines.

| Cluster type | Partial type | Sensitivity | 1-Specificity | Precision | F1 |
|---|---|---|---|---|---|
| Histopathology | 0 | 0.670 (0.036) | 0.330 (0.036) | 0.739 (0.015) | 0.639 (0.053) |
| Histopathology | 10 | 0.622 (0.034) | 0.378 (0.034) | 0.709 (0.022) | 0.570 (0.053) |
| Histopathology | 20 | 0.607 (0.010) | 0.393 (0.010) | 0.715 (0.007) | 0.546 (0.016) |
| Histopathology | 30 | 0.592 (0.027) | 0.408 (0.027) | 0.723 (0.015) | 0.515 (0.045) |
| Histopathology | 40 | 0.571 (0.015) | 0.429 (0.015) | 0.752 (0.022) | 0.471 (0.031) |
| Histopathology | 50 | 0.541 (0.016) | 0.459 (0.016) | 0.724 (0.056) | 0.417 (0.031) |
| Histopathology | Natural | 0.702 (0.047) | 0.298 (0.047) | 0.758 (0.031) | 0.679 (0.057) |
| ImageNet | 0 | 0.644 (0.033) | 0.356 (0.033) | 0.726 (0.021) | 0.603 (0.047) |
| ImageNet | 10 | 0.609 (0.023) | 0.391 (0.023) | 0.704 (0.022) | 0.551 (0.040) |
| ImageNet | 20 | 0.626 (0.033) | 0.374 (0.033) | 0.732 (0.032) | 0.570 (0.047) |
| ImageNet | 30 | 0.586 (0.026) | 0.414 (0.026) | 0.722 (0.024) | 0.504 (0.050) |
| ImageNet | 40 | 0.563 (0.018) | 0.437 (0.018) | 0.725 (0.017) | 0.462 (0.038) |
| ImageNet | 50 | 0.564 (0.028) | 0.436 (0.028) | 0.744 (0.025) | 0.457 (0.052) |
| ImageNet | Natural | 0.644 (0.020) | 0.356 (0.020) | 0.725 (0.025) | 0.604 (0.028) |
| Single | 0 | 0.677 (0.039) | 0.323 (0.039) | 0.742 (0.022) | 0.648 (0.057) |
| Single | 10 | 0.662 (0.017) | 0.338 (0.017) | 0.742 (0.017) | 0.626 (0.022) |
| Single | 20 | 0.622 (0.028) | 0.378 (0.028) | 0.731 (0.030) | 0.565 (0.038) |
| Single | 30 | 0.622 (0.017) | 0.378 (0.017) | 0.746 (0.023) | 0.562 (0.026) |
| Single | 40 | 0.597 (0.030) | 0.403 (0.030) | 0.746 (0.020) | 0.518 (0.057) |
| Single | 50 | 0.555 (0.024) | 0.445 (0.024) | 0.751 (0.019) | 0.439 (0.045) |
| Single | Natural | **0.704 (0.045)** | **0.296 (0.045)** | **0.767 (0.031)** | **0.680 (0.056)** |

**Table 5**. LNM classification results for all individual models when positive/negative classes are considered. Sensitivity, 1-Specificity, Precision, and F1 are reported as mean (SD) over 5-folds. Top performers are in bold.

| Ensemble name | Sensitivity | 1-Specificity | Precision | F1 |
|---|---|---|---|---|
| Generalizing | 0.765 (0.011) | 0.235 (0.011) | 0.787 (0.009) | 0.758 (0.012) |
| Non-Generalizing (8) | 0.706 (0.010) | 0.294 (0.010) | 0.749 (0.010) | 0.689 (0.012) |
| Generalizing (T&N) | **0.806 (0.024)** | **0.194 (0.024)** | **0.809 (0.020)** | **0.805 (0.025)** |
| Generalizing (T) | 0.719 (0.028) | 0.281 (0.028) | 0.763 (0.022) | 0.703 (0.032) |
| Generalizing (N) | 0.762 (0.012) | 0.238 (0.012) | 0.788 (0.009) | 0.753 (0.013) |
| Natural% | 0.745 (0.027) | 0.255 (0.027) | 0.778 (0.025) | 0.734 (0.029) |
| 0% | 0.739 (0.026) | 0.261 (0.026) | 0.777 (0.020) | 0.726 (0.030) |
| 10% | 0.678 (0.025) | 0.322 (0.025) | 0.738 (0.015) | 0.652 (0.034) |
| 20% | 0.662 (0.022) | 0.338 (0.022) | 0.742 (0.019) | 0.626 (0.028) |
| 30% | 0.659 (0.018) | 0.341 (0.018) | 0.756 (0.020) | 0.618 (0.026) |
| 40% | 0.618 (0.019) | 0.382 (0.019) | 0.746 (0.012) | 0.554 (0.031) |
| 50% | 0.577 (0.016) | 0.423 (0.016) | 0.717 (0.029) | 0.489 (0.024) |
| Histopathology | 0.686 (0.015) | 0.314 (0.015) | 0.739 (0.017) | 0.664 (0.017) |
| ImageNet | 0.668 (0.011) | 0.332 (0.011) | 0.731 (0.012) | 0.640 (0.014) |
| Single | 0.724 (0.022) | 0.276 (0.022) | 0.778 (0.021) | 0.706 (0.025) |

**Table 6**. LNM classification results for all ensemble models when positive/negative classes are considered. Sensitivity, 1-Specificity, Precision, and F1 are reported as mean (SD) over 5-folds. Top performers are in bold.

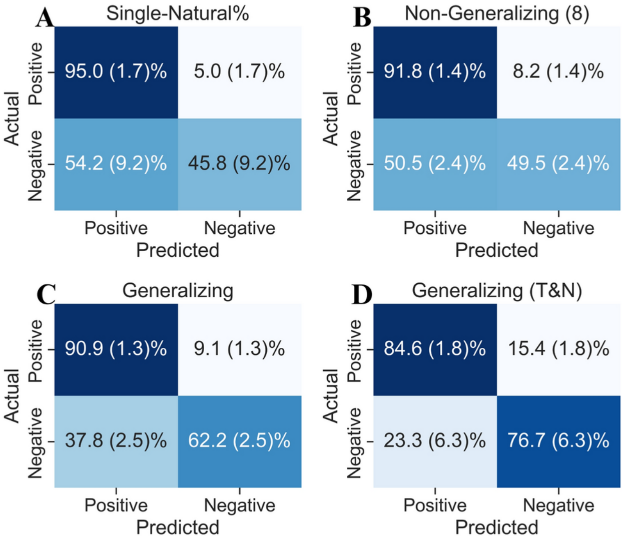

**Fig. 8**. Confusion matrices for LNM classification of the positive (macro/micro) and negative (ITC/negative) classes reported as mean (SD) percentages of the actual class. (**A**) Single-Natural%. (**B**) Non-Generalizing (8). (**C**) Generalizing. (**D**) Generalizing (T&N).

and increased FPR for patch-level segmentation with ImageNet models. For normal patches, both methods have 5 clusters for normal-normal and 6 clusters for normal-tumor patch groups and appear to be able to separately cluster patches with a high density of lymphocytes and those with adipose tissue. However, Histopathology has additional clusters for patches with both lymphocytes and histiocytes, which is unseen with ImageNet. Also, Histopathology can group blurry patches that could be problematic for false positives (which are lacking in a typical dataset). On the other hand, ImageNet looks to have a dedicated cluster to stroma, which could potentially improve robustness for this specific tissue type.

It is interesting to note that one of the two models that had both tumor and normal generalization was Single-Natural%, with the other being Histopathology-0%. The performance of Single-Natural% and Histopathology-0% on their own was lower than the generalizing ensemble (T&N) with these two models. This suggests that while Single-Natural% may capture some of the natural variability in the datasets due to random sampling, the histopathology model was necessary to add the details that perhaps are more specific for LNM segmentation. Furthermore, the ImageNet ensemble performed poorly for all metrics, which underscores the need for more foundation and domain-specific pre-trained models specifically for pathology.

The PT played an important role in the performance across the LNM category. A higher PT percentage resulted in improved tumor segmentation particularly in ITCs at the expense of increased false positives. This is likely due to the fact that tumor boundaries are used in the training dataset and since ITCs are a few patches, and mainly composed of patches with only PT, this improves performance. However, since the variability of the appearance of boundaries (and percentage of PT) is vast, it is recommended that future studies include more PT in the training set to account for the variability. We find this to be consistent with existing class imbalance literature[50–52] that discusses how learners often show a preference for the majority class, at the the cost of overlooking the minority class. While the full tumor and full normal patches (0% PT) had reduced performance in the smaller objects, it offered generalization in both tumor and normal patches for the histopathology models and improved performance in the final WSI classification results.

While we were able to balance datasets and focus on models that generalize, we are aware that of the limitation of current models in that we can only reach a certain accuracy due to the tradeoff between TP and FP. Future works should focus more on the FP rate and we suspect the FP rate from patch-level predictions in digital pathology studies is largely under-reported. This could be a significant challenge if reporting requirements for ITC change. Currently, ITC are acknowledged if detected during clinical reporting, but the TNM staging system[47] still categorizes the lymph node as negative if there are no macro or micro metastases found. The clinical significance of ITC and micro metastases in lymph nodes remains a matter of debate[5,53,54], although after preoperative treatment, the relevance of any size of residual metastasis in the lymph node comes into consideration[48,55]. Improving ITC performance is crucial and could be achieved with more boundary patches and methods that reduce FP rates.

Limitations of this study include sample size, ITC availability, and annotation quality. The inclusion of more ID and OOD data samples for each LNM would enhance the robustness of the findings. This study is limited by the training dataset size of 35, 000 patches due to the number of PT patches. In the future, larger sample sizes could improve generalization and overall performance. Additionally, the choice of equivalence interval should be validated against domain specific knowledge. Generalizability of the results depends on the datasets used for ID and OOD data. Therefore, ensuring that these datasets are representative of real-world variability is crucial for the external validity of the findings. It should also be noted that the power of TOST depends on the sample size. With 21 models, the study might have limited power to detect small but clinically meaningful differences. On the note of ITC availability, the models were not trained on ITC patches since there were no ITC slides available in CAMELYON16. While we propose methods to overcome this through more inclusion of boundary patches, this challenge has also been acknowledged by other authors[5,56], proving to be an unsolved problem. Lastly, any inaccuracy in annotations can create noise in the results. For example, annotations of large tumor masses including regions of normal cells can inflate the FPR, as illustrated in Supplementary Fig. S6. This issue is compounded in lobular carcinoma, where tumorous regions are dispersed.

Several avenues could be explored in the future. Improving the segmentation of ITC is crucial, and adding ITC patches to the training set could be beneficial due to the unique morphology they may present. This may still not be enough to achieve the desired result due to the level of difficulty, so the adoption of recent state-of-the-art architectures like Vision Transformers[57] and ConvNeXt[58] may offer considerable improvements as well. We also plan to explore segmentation models such as U-Net[59] for their capability to extract tissue regions compared to Otsu for foreground segmentation as was done in this work. To mitigate the heavy reliance on extensively annotated WSI which are difficult to obtain, self-supervised[36,60–62] or generative augmentation[63–65] methods can be considered. Additionally, experimenting with alternative stain normalization methods[66–68] or stain augmentation techniques[69–72] could improve generalization. The lack of contextual information in the patch size-magnification combination used in this study ($256 \times 256$, $20\times$) can lead to segmentation errors on small objects. We must also consider that pathologists utilize surrounding areas and multiple magnifications to come to their conclusions. Therefore, incorporating multi-resolution architectures[62,73–78] could provide the necessary context for more accurate segmentations. Finally, the method for determining the optimal number of clusters may result in cluster overlap. Future work will focus on addressing this issue by exploring alternative clustering algorithms[79,80] or refining selection criteria to minimize overlap and enhance cluster distinctiveness.

Pathology provides the definitive diagnosis, and AI tools are poised to improve accuracy, inter-rater agreement, and turn-around time (TAT) of pathologists, leading to improved quality of care[81]. A fundamental challenge of implementing AI tools widely is OOD generalization, where performance in data not used in the training set is drastically reduced due to domain shift. For the task of LNM classification, recent works found that on OOD data, there was a drop in classification performance which necessitated retraining on the OOD data to improve performance[48]. Re-training algorithms on OOD data for pathology imaging is impractical and new methods are needed to overcome this key barrier. Therefore, this work proposes methods to generate training datasets in an unsupervised manner that generalize better to OOD data, as well as a method to measure the reliability and generalization in ID and OOD data. Ensuring the performance across datasets is equivalent is necessary for optimized quality of care and patient safety.

## Conclusions

This study evaluates models constructed from various cluster-based sampling techniques in terms of their performance in LNM segmentation at both the patch and WSI levels, as well as LNM classification using an explainable and clinically informed decision structure. Notably, our ensemble models, selected based on a novel generalization framework, consistently outperformed individual models. This framework determines models that demonstrate robust generalization across both tumor and normal tissues, enhancing performance in WSI-level segmentation and LNM classification when ensembled. This study highlights the potential of curated training sets and our proposed generalization framework to refine ensemble techniques, thus enhancing clinical utility and improving patient care.

## Data availability

The open-source datasets used in this study (CAMELYON16 and CAMELYON17) are available at https://camelyon17.grand-challenge.org/ Data/ through GigaScience, AWS or Baidu channels. Private clinical data is not available. Segmentation models are available at https://github.com/IAMLAB-Ryerson/OOD-Generalization-LNM.

Received: 28 June 2024; Accepted: 19 November 2024
Published online: 07 January 2025

## References

1. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **74**, 12–49. https://doi.org/10.3322/caac.21820 (2024).
2. Budginaite, E., Magee, D. R., Kloft, M., Woodruff, H. C. & Grabsch, H. I. Computational methods for metastasis detection in lymph nodes and characterization of the metastasis-free lymph node microarchitecture: A systematic-narrative hybrid review. *J. Pathol. Inform.* **15**, 100367. https://doi.org/10.1016/j.jpi.2024.100367 (2024).
3. Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F. & Park, J. Y. Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Netw. Open* **2**, e194337. https://doi.org/10.1001/jamanetworkopen.2019.4337 (2019).
4. Cohen, M. B. et al. Features of burnout amongst pathologists: A reassessment. *Acad. Pathol.* **9**, 100052. https://doi.org/10.1016/j.acpath.2022.100052 (2022).
5. Bándi, P. et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Trans. Med. Imaging.* **38**, 550–560. https://doi.org/10.1109/TMI.2018.2867350 (2019).
6. Litjens, G. et al. 1399 H &E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience.* **7**, giy065. https://doi.org/10.1093/gigascience/giy065 (2018).
7. EhteshamiBejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* **318**, 2199–2210. https://doi.org/10.1001/jama.2017.14585 (2017).
8. Deng, S. et al. Deep learning in digital pathology image analysis: a survey. *Front. Med.* **14**, 470–487. https://doi.org/10.1007/s11684-020-0782-9 (2020).
9. Wang, L. et al. Deep regional metastases segmentation for patient-level lymph node status classification. *IEEE Access.* **9**, 129293–129302. https://doi.org/10.1109/ACCESS.2021.3113036 (2021) (**Conference Name: IEEE Access.**).
10. Kleppe, A. et al. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer.* **21**, 199–211. https://doi.org/10.1038/s41568-020-00327-9 (2021) (**Publisher: Nature Publishing Group.**).
11. Scalbert, M., Vakalopoulou, M. & Couzinie-Devy, F. Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology. https://doi.org/10.48550/arXiv.2206.09769 (2022).
12. Razavi, S. et al. MiNuGAN: Dual segmentation of mitoses and nuclei using conditional GANs on multi-center breast H &E images. *J. Pathol. Inform.* **13**, 100002. https://doi.org/10.1016/j.jpi.2022.100002 (2022).
13. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G. & Srinivasan, B. A generalized deep learning framework for whole-slide image segmentation and analysis. *Sci. Rep.* **11**, 11579. https://doi.org/10.1038/s41598-021-90444-8 (2021) (**Number: 1 Publisher: Nature Publishing Group.**).
14. Zhong, A. & Li, Q. Team HMS-MGH-CCDS method1 (2018).
15. Zanjani, F. G., Zinger, S. & de, P. H. N. Automated detection and classification of cancer metastases in whole-slide histopathology images using deep learning (2017).
16. Fukuta, K., Komura, D., Harada, T. & Ishikawa, S. Identifying metastatic breast cancer using deep texture representation (2017).
17. Taori, R. *et al.* Measuring robustness to natural distribution shifts in image classification. in *Advances in Neural Information Processing Systems*, vol. 33, 18583–18599 (Curran Associates, Inc., 2020).
18. Andreassen, A., Bahri, Y., Neyshabur, B. & Roelofs, R. The evolution of out-of-distribution robustness throughout fine-tuning. https://doi.org/10.48550/arXiv.2106.15831 (2021).
19. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. Do ImageNet classifiers generalize to ImageNet? in *Proceedings of the 36th International Conference on Machine Learning*, 5389–5400 (PMLR, 2019). ISSN: 2640-3498.
20. Wenzel, F. et al. Assaying out-of-distribution generalization in transfer learning. *Adv. Neural Inform. Process. Syst.* **35**, 7181–7198 (2022).
21. Teney, D., Lin, Y., Oh, S. J. & Abbasnejad, E. ID and OOD performance are sometimes inversely correlated on real-world datasets. https://doi.org/10.48550/arXiv.2209.00613 (2023).
22. Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinetics Biopharm.* **15**, 657–680. https://doi.org/10.1007/BF01068419 (1987).
23. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 [cs] (2017).
24. Pathcore. PathcoreFlow (Version 3.1.1) (2024).
25. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27. https://doi.org/10.4103/2153-3539.119005 (2013).
26. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250 (2009). ISSN: 1945-8452.
27. Pocock, J. et al. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Commun. Med.* **2**, 1–14. https://doi.org/10.1038/s43856-022-00186-5 (2022) (**Number: 1 Publisher: Nature Publishing Group.**).
28. Khan, A. et al. Impact of scanner variability on lymph node segmentation in computational pathology. *J. Pathol. Inform.* **13**, 100127. https://doi.org/10.1016/j.jpi.2022.100127 (2022).
29. Pontalba, J. T. *et al.* Assessing the impact of color normalization in convolutional neural network-based nuclei segmentation frameworks. *Front. Bioeng. Biotechnol.* **7** (2019).
30. Li, W. & Chen, W. Reproducibility in deep learning algorithms for digital pathology applications: A case study using the CAMELYON16 datasets. in *Medical Imaging 2021: Digital Pathology*, vol. 11603, 323–332. https://doi.org/10.1117/12.2581996 (SPIE, 2021).
31. Bándi, P., Balkenhol, M., van Ginneken, B., van der Laak, J. & Litjens, G. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. *PeerJ* **7**, e8242. https://doi.org/10.7717/peerj.8242 (2019).
32. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybernetics* **9**, 62–66. https://doi.org/10.1109/TSMC.1979.4310076 (1979) (**Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.**).
33. Sikaroudi, M., Hosseini, M., Gonzalez, R., Rahnamayan, S. & Tizhoosh, H. R. Generalization of vision pre-trained models for histopathology. *Sci. Rep.* **13**, 6065. https://doi.org/10.1038/s41598-023-33348-z (2023) (**Number: 1 Publisher: Nature Publishing Group.**).
34. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90 (2016). (**ISSN: 1063-6919**).
35. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848 (2009). (**ISSN: 1063-6919**).

36. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* **7**, 100198. https://doi.org/10.1016/j.mlwa.2021.100198 (2022).
37. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276. https://doi.org/10.1007/BF02289263 (1953).
38. Sammouda, R. & El-Zaart, A. An optimized approach for prostate image segmentation using k-means clustering algorithm with elbow method. *Comput. Intell. Neurosci.* **2021**, 4553832. https://doi.org/10.1155/2021/4553832 (2021).
39. Shi, C. et al. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J. Wireless Commun. Netw.* **2021**, 1–16. https://doi.org/10.1186/s13638-021-01910-w (2021) (**Number: 1 Publisher: SpringerOpen.**).
40. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27. https://doi.org/10.1080/03610927408827101 (1974) (**Publisher: Taylor & Francis**).
41. Davies, D. L. & Bouldin, D. W. A Cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI–1**, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909 (1979) (**Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.**).
42. Howard, A. *et al.* Searching for MobileNetV3, https://doi.org/10.48550/arXiv.1905.02244 (2019).
43. Abraham, N. & Khan, N. M. A Novel focal Tversky loss function with improved attention U-Net for lesion segmentation. in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 683–687. https://doi.org/10.1109/ISBI.2019.8759329 (2019). (**ISSN: 1945-8452**).
44. Tversky, A. Features of similarity. *Psychol. Rev.* **84**, 327–352. https://doi.org/10.1037/0033-295X.84.4.327 (1977) (**Place: US Publisher: American Psychological Association.**).
45. Caldwell, A. R. Exploring equivalence testing with the updated TOSTER R package. Preprint, PsyArXiv (2022). https://doi.org/10.31234/osf.io/ty8de.
46. Lakens, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychol. Personality Sci.* **8**, 355–362. https://doi.org/10.1177/1948550617697177 (2017).
47. Brierley, J., Gospodarowicz, M. K. & Wittekind, C. (eds.) *TNM Classification of Malignant Tumours*, 8h edn. (Wiley Blackwell, 2017).
48. Jarkman, S. et al. Generalization of deep learning in digital pathology: Experience in breast cancer metastasis detection. *Cancers* **14**, 5424. https://doi.org/10.3390/cancers14215424 (2022) (**Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.**).
49. Edge, S. B. & American Joint Committee on Cancer (eds.) *AJCC Cancer Staging Manual* (Springer, 2010), 7th edn. OCLC: ocn316431417.
50. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27. https://doi.org/10.1186/s40537-019-0192-5 (2019).
51. Khan, A. A., Chaudhari, O. & Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* **244**, 122778. https://doi.org/10.1016/j.eswa.2023.122778 (2024).
52. Rezvani, S. & Wang, X. A broad review on class imbalance learning techniques. *Appl. Soft Comput.* **143**, 110415. https://doi.org/10.1016/j.asoc.2023.110415 (2023).
53. Houvenaeghel, G. et al. Lack of prognostic impact of sentinel node micro-metastases in endocrine receptor-positive early breast cancer: Results from a large multicenter cohort. *ESMO Open* **6**, 100151. https://doi.org/10.1016/j.esmoop.2021.100151 (2021).
54. Apple, S. K. Sentinel lymph node in breast cancer: Review article from a pathologist's point of view. *J. Pathol. Transl. Med.* **50**, 83–95. https://doi.org/10.4132/jptm.2015.11.23 (2016).
55. Viale, G. & Fusco, N. Pathology after neoadjuvant treatment—How to assess residual disease. *Breast (Edinburgh, Scotland)* **62**(Suppl 1), S25–S28. https://doi.org/10.1016/j.breast.2021.11.009 (2022).
56. Khalil, M.-A., Lee, Y.-C., Lien, H.-C., Jeng, Y.-M. & Wang, C.-W. Fast segmentation of metastatic foci in H &E whole-slide images for breast cancer diagnosis. *Diagnostics* **12**, 990. https://doi.org/10.3390/diagnostics12040990 (2022) (**Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.**).
57. Dosovitskiy, A. *et al.* An image is worth 16 x 16 words: Transformers for image recognition at scale. arXiv:2010.11929 [cs]
58. Liu, Z. *et al.* A ConvNet for the 2020s, https://doi.org/10.48550/arXiv.2201.03545 (2022). ArXiv:2201.03545 [cs].
59. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. in (Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F., eds.) *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Lecture Notes in Computer Science*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28 (Springer International Publishing, 2015).
60. Srinidhi, C. L., Kim, S. W., Chen, F.-D. & Martel, A. L. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med. Image Anal.* **75**, 102256. https://doi.org/10.1016/j.media.2021.102256 (2022).
61. Wang, X. *et al.* TransPath: Transformer-based self-supervised learning for histopathological image classification. in (de Bruijne, M. et al., eds.) *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Lecture Notes in Computer Science*, 186–195. https://doi.org/10.1007/978-3-030-87237-3_18 (Springer International Publishing, 2021).
62. Chen, R. J. *et al.* Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. https://doi.org/10.48550/arXiv.2206.02647 (2022).
63. Jose, L., Liu, S., Russo, C., Nadort, A. & Di Ieva, A. Generative adversarial networks in digital pathology and histopathological image processing: A review. *J. Pathol. Inform.* **12**, 43. https://doi.org/10.4103/jpi.jpi_103_20 (2021).
64. Vasiljević, J., Feuerhake, F., Wemmert, C. & Lampert, T. Towards histopathological stain invariance by Unsupervised Domain Augmentation using generative adversarial networks. *Neurocomputing* **460**, 277–291. https://doi.org/10.1016/j.neucom.2021.07.005 (2021).
65. Wei, J. et al. Generative image translation for data augmentation in colorectal histopathology images. *Proc. Mach. Learn. Res.* **116**, 10–24 (2019).
66. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41. https://doi.org/10.1109/38.946629 (2001) (**Conference Name: IEEE Computer Graphics and Applications.**).
67. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971. https://doi.org/10.1109/TMI.2016.2529665 (2016).
68. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Analyt. Quant. Cytol. Histol.* **23**, 291–299 (2001).
69. Chang, J.-R. *et al.* Stain mix-up: Unsupervised domain generalization for histopathology images. in (de Bruijne, M. et al., eds.) *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Lecture Notes in Computer Science*, 117–126. https://doi.org/10.1007/978-3-030-87199-4_11 (Springer International Publishing, 2021).
70. Cho, H., Lim, S., Choi, G. & Min, H. Neural stain-style transfer learning using GAN for histopathological images. https://doi.org/10.48550/arXiv.1710.08543 (2017).
71. Zanjani, F. G., Zinger, S., Bejnordi, B. E., van der Laak, J. A. W. M. & de With, P. H. N. Stain normalization of histopathology images using generative adversarial networks. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 573–577. https://doi.org/10.1109/ISBI.2018.8363641 (2018). (**ISSN: 1945-8452**).
72. Bug, D. *et al.* Context-based normalization of histological stains using deep convolutional features. in (Cardoso, M. J. et al., eds.) *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Lecture Notes in Computer Science*, 135–142. https://doi.org/10.1007/978-3-319-67558-9_16 (Springer International Publishing, 2017).

73. Deng, R. et al. Cross-scale multi-instance learning for pathological image diagnosis. *Med. Image Anal.* **94**, 103124. https://doi.org/10.1016/j.media.2024.103124 (2024).
74. Das, R., Bose, S., Chowdhury, R. S. & Maulik, U. Dense dilated multi-scale supervised attention-guided network for histopathology image segmentation. *Comput. Biol. Med.* **163**, 107182. https://doi.org/10.1016/j.compbiomed.2023.107182 (2023).
75. Ning, Z. et al. Mutual-assistance learning for standalone mono-modality survival analysis of human cancers. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 7577–7594. https://doi.org/10.1109/TPAMI.2022.3222732 (2023) (**Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.**).
76. Ding, Y. et al. Multi-center study on predicting breast cancer lymph node status from core needle biopsy specimens using multi-modal and multi-instance deep learning. *NPJ Breast Cancer.* **9**, 1–13. https://doi.org/10.1038/s41523-023-00562-x (2023) (**Publisher: Nature Publishing Group.**).
77. Ho, D. J. et al. Deep multi-magnification networks for multi-class breast cancer image segmentation. *Comput. Med. Imaging Graph.* **88**, 101866. https://doi.org/10.1016/j.compmedimag.2021.101866 (2021).
78. Tu, C., Zhang, Y. & Ning, Z. Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images. *Adv. Neural Inform. Process. Syst.* **35**, 29484–29497 (2022).
79. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, 226–231 (AAAI Press, 1996).
80. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM SIGMOD Rec.* **28**, 49–60. https://doi.org/10.1145/304181.304187 (1999).
81. Dy, A. et al. AI improves accuracy, agreement and efficiency of pathologists for Ki67 assessments in breast cancer. *Sci. Rep.* **14**, 1283. https://doi.org/10.1038/s41598-024-51723-2 (2024) (**Publisher: Nature Publishing Group.**).

## Acknowledgements

This research acknowledges the Canada Research Chairs Program, Canadian Cancer Society, and Mitacs.

## Author contributions

Y.V. designed the study, performed all data collection and pre-processing, implemented, trained, and evaluated all models, designed and implemented the proposed statistical method, interpreted the data, created all tables and figures. P.T. designed the proposed statistical method. K.J. annotated the clinical dataset. A.K. (principal investigator) designed the study, contributed to the interpretation of the data, analyzed the results, helped draft the manuscript, and supervised the project. All authors reviewed the manuscript.

## Funding

This research is funded by the Canada Research Chairs Program, Canadian Cancer Society (Emerging Scholars) and Mitacs (Accelerate).

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80495-y.

**Correspondence** and requests for materials should be addressed to Y.V.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.