scientific reports

Check for updates

OPEN

# Leveraging immuno-fluorescence data to reduce pathologist annotation requirements in lung tumor segmentation using deep learning

Hatef Mehrabian✉, Jens Brodbeck, Peipei Lyu, Edith Vaquero, Abhishek Aggarwal & Lauri Diehl

**The main bottleneck in training a robust tumor segmentation algorithm for non-small cell lung cancer (NSCLC) on H&E is generating sufficient ground truth annotations. Various approaches for generating tumor labels to train a tumor segmentation model was explored. A large dataset of low-cost low-accuracy panCK-based annotations was used to pre-train the model and determine the minimum required size of the expensive but highly accurate pathologist annotations dataset. PanCK pre-training was compared to foundation models and various architectures were explored for model backbone. Proper study design and sample procurement for training a generalizable model that captured variations in NSCLC H&E was studied. H&E imaging was performed on 112 samples (three centers, two scanner types, different staining and imaging protocols). Attention U-Net architecture was trained using the large panCK-based annotations dataset (68 samples, total area 10,326 [mm$^2$]) followed by fine-tuning using a small pathologist annotations dataset (80 samples, total area 246 [mm$^2$]). This approach resulted in mean intersection over union (mIoU) of 82% [77 87]. Using panCK pretraining provided better performance compared to foundation models and allowed for 70% reduction in pathologist annotations with no drop in performance. Study design ensured model generalizability over variations on H&E where performance was consistent across centers, scanners, and subtypes.**

**Keywords** Non-small cell lung cancer (NSCLC), Tumor segmentation, Convolutional neural network (CNN), Digital pathology, panCK tumor annotation

Lung cancer is the leading cause of cancer-related deaths worldwide, with non-small cell lung cancer (NSCLC) accounting for approximately 85% of all cases[1,2]. Pathologist review of hematoxylin and eosin (H&E) stained tissue sections, immuno-histochemistry (IHC), and molecular data are the gold standard for NSCLC diagnosis and treatment planning[3]. Pathologists rely on spatial information present in H&E images, such as tissue and cell morphology, nuclear atypia, and other relevant features, to assess tumors and their progression[4]. Such analysis is typically performed manually by the pathologists with significant interobserver variability[5].

NSCLC includes three main subtypes: adenocarcinoma (LUAD, 40%) and squamous cell carcinoma (LUSC, 25%) and large cell carcinoma (LCC, 10–15%)[6]. According to world health organization (WHO), the principles for lung cancer classification is morphology followed by IHC and then molecular techniques, which directly impact patient management and treatment options[6–8]. NSCLC subtypes have distinct histological and morphological characteristics on H&E images[6,7], for instance LUAD typically shows glandular differentiation patterns such as lepidic, acinar, and papillary growth patterns[9], while LUSC are often composed of large keratinizing cells with intercellular bridges[10]. Prognostic importance of nuclear morphology as seen on H&E have been demonstrated in several studies[11] where malignant cells often manifest with hyperchromatic nuclei with irregular shapes[12]. The membranes of these malignant cells also tend to be different in shape, potentially due to altered cell division. In addition to morphological features, IHC characteristics also plays a crucial role in classification of NSCLC, where LUAD and LUSC can be efficiently separated using thyroid transcription factor 1 (TIF1) and p40 staining[13].

Non-Clinical Safety and Pathobiology, Gilead Sciences, Foster City, CA, USA. ✉email: hatef.mehrabian@gilead.com

A comprehensive tissue analysis often requires considering H&E and IHC or immuno-fluorescent (IF) images together, and tumor segmentation is often an essential first step in such analysis. Tumor can be annotated in H&E whole slide images (WSI), however it is not practical for a pathologist to manually annotate multiple WSI. Thus, automated tumor segmentation is required for a compartmentalized image analysis that separates tumor from surrounding tissues. Such tumor annotation can then be transferred to the IHC or IF image for further marker analysis.

Recent advances in deep learning have led to the development of convolutional neural networks (CNNs) that can effectively segment tumor regions on histological images of NSCLC[14–20]. The main bottleneck in developing a robust model is lack of sufficiently large annotated datasets to capture all NSCLC variations on H&E WSI. The gold standard is manual pathologist annotation, which can only be generated on a small dataset due to being expensive and time-consuming. However, the variability in NSCLC subtypes, as well as staining protocols, scanners, operators, and centers pose challenges for developing generalizable segmentation models[21–23] by increasing the size of the required manual annotations. Thus, techniques that reduce the reliance on such manual annotations is desired. Using IHC or IF markers such as Pan Cytokeratin (panCk) to generate a large amount of tumor annotations is promising but suffers from non-specificness and technical challenges. The objective of this study was to investigate different approaches for generating ground truth annotations and to quantify the impact of each approach and their combinations in the final model performance, and in determining the minimum amount of required pathologist annotations. Moreover, different pre-training approaches such as using foundation models[24–26] and the impact of different model and backbone architectures[27–30] were investigated. However, the main aim of this study was investigating the less studied aspect of training a segmentation model, which is tumor annotation generation techniques and determining best practices for such data curation. Here, a hybrid model that leveraged a large dataset annotated with panCK, combined with a small dataset annotated manually by pathologists was investigated. A comprehensive study of the impacts of study design factors on model accuracy and how much reduction in manual annotation is feasible without impacting performance was also performed to determine the optimal workflow with minimal manual annotations.

There exist a rich body of work in lung tumor segmentation on H&E. Tokunaga et al.[14] proposed a U-Net-based method with adaptive weighting for different acquisition magnifications for NSCLC segmentation in H&E images (on a dataset of 29 WSI), achieving mean intersection over union (mIoU) of 83%, compared to a single resolution U-Net at 20x magnification with mIoU of 77%. Arlova et al.[19] used 239 manually annotated mouse lung WSIs and trained segmentation models using U-Net and DeepLabV3 + with backbones from Resnet18, Resnet34, ResNet50. They achieved mIoU scores ranging from 76% to 80% for various combinations of the model architectures and backbones. Raczkowski et al.[15] used a ResNet-inspired network for lung tumor microenvironment segmentation and used it to predict tumor mutation and patient survival. Wei et al.[16] used a ResNet18-based model to classify lung adenocarcinoma patterns for patient prognosis and survival evaluation and demonstrated similar performance to expert pathologists.

There have also been two main competitions for lung tumor segmentation on H&E WSI, the ACDC@lungHP challenge in 2019[17] and the WSSS4LUAD challenge in 2021[18], using tumor annotations generated manually by pathologists. The ACDC@lungHP challenge focused on the segmentation of both lung adenocarcinomas (LUAD) and lung squamous cell carcinomas (LUSC) using supervised learning techniques and provided pixel-level annotations on a dataset of 150 NCSLC WSI that were scanned using a single digital scanner (3DHISTECH Panoramic 250) and annotated by one pathologist with 30 years of experience. The top 10 submitted models provided an average pixel-level Dice coefficient of 80% for LUSC and 77% for LUAD (best model Dice coefficient = 84%). On the other hand, the WSSS4LUAD challenge focused on weakly supervised segmentation, where patch-level annotations of H&E images were provided on a dataset of 63 LUAD WSI only. The top 10 models provided an average patch-level mIoU of 79% (best model mIoU = 84%).

These competitions have spurred significant research efforts in developing accurate and efficient methods for lung tumor segmentation. These challenges demonstrated the potential of machine learning methods for accurately segmenting tumors in H&E images, even with limited annotations. However, these competitions also highlight the need for continued research to improve the accuracy of tumor segmentation, particularly for the challenging LUSC subtype[17,18].

PanCk is a sensitive marker for NSCLC tumor tissue and previous studies have shown that panCk-based annotations can be used to train CNNs for tumor segmentation[31–33]. PanCK has also been used in generating tumor annotation for training tumor segmentation models for prostate[34] and breast[35] cancers. However, panCk expression is not specific to the tumor tissue only. It is also expressed in non-tumor tissues such as necrotic tissue, normal lung, and auto-fluorescence from red blood cells which is often captured in panCK imaging channel (Fig. 1a). Thus, the use of panCk-based tumor annotations results in false positives and reduces the overall accuracy of the segmentation model. Weis et al.[31] used IHC images of panCk to generate ground truth tumor annotation on a serial section to the H&E tissue micro-arrays (TMA) from 247 samples and trained a U-Net model, achieving an average pixel-level Dice coefficient of 44% and accuracy of 70% (using balanced cross-entropy loss function and $512 \times 512$ patch size). Kapil et al.[32] also used panCK annotations for tumor segmentation with a goal of cell scoring and survival analysis on PD-L1 images and reported F1 score of 55% for tumor segmentation.

In addition to the limited availability of annotated data, the variability in staining protocols, scanners, operators, and centers can also pose challenges for developing accurate segmentation models[21–23]. Color variations occur in hematoxylin and eosin stains between studies, and even when using the same protocol (depending on the tissue and slide preparations procedures). To address these challenges, multiple studies have proposed different strategies, including the use of data and color augmentation[19,23,36], domain adaptation[32,37], and multi-center datasets[38]. These approaches ensure the developed model is robust and generalizable across centers and protocols.

Considering various NSCLC subtypes is also important in the accurate segmentation of tumors in H&E images, given their different morphological features. The ACDC@lungHP competition showed that model

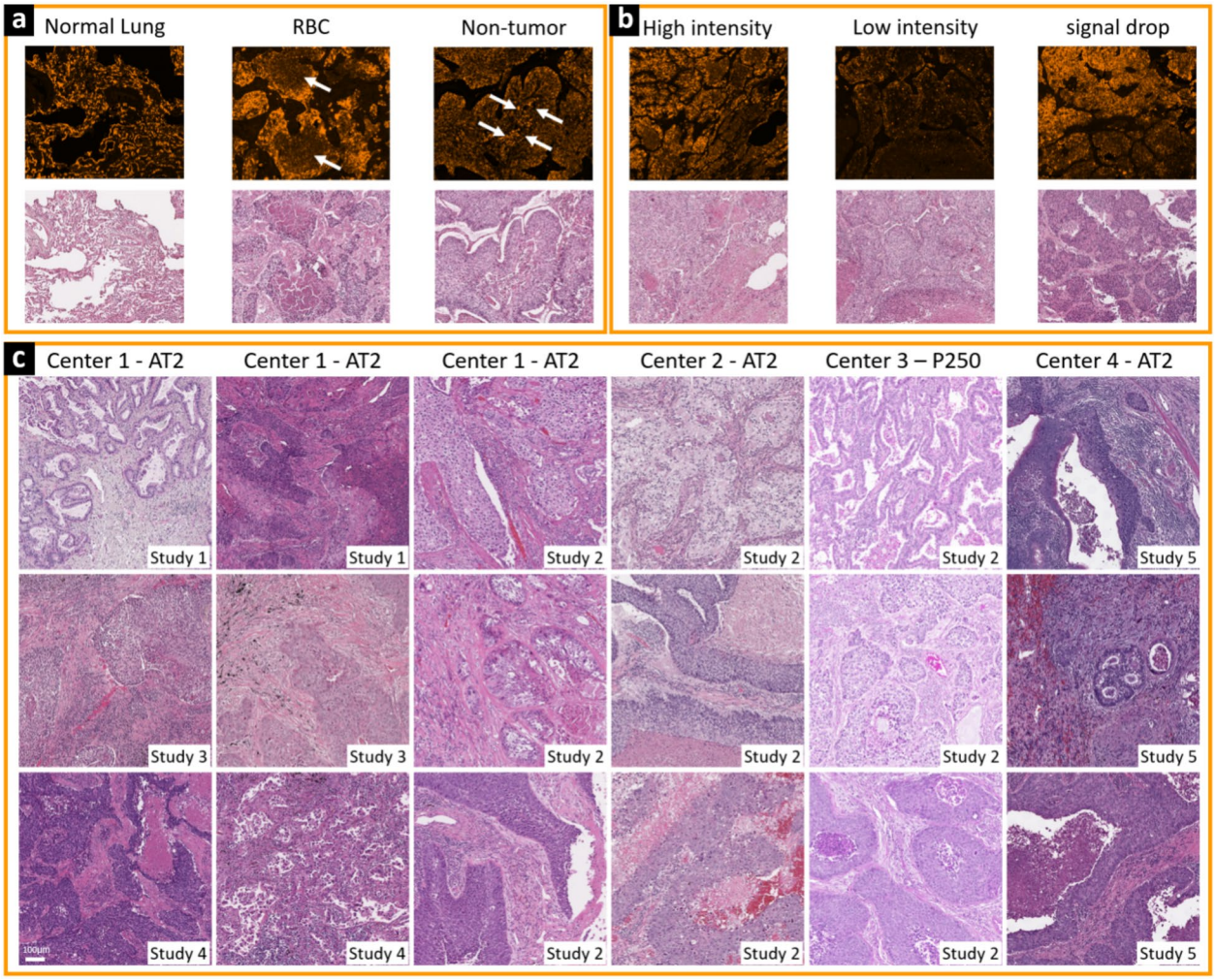


**Figure 1.** Examples of panCK expression in non-tumor tissues (**a**), Technical difficulties in acquiring panCK images (**b**), color and morphological variation in NSCLC tumor tissue due to differences in center, protocol, scanner, and operator (**c**). Study 5 shows images from a study that was not involved in data preparation (included here to demonstrate extent of color variaiton).

performance varied across different NSCLC subtypes[17]. Some studies have addressed this challenge by incorporating multi-scale analysis and prior knowledge of the morphological features of different NSCLC subtypes[39].

In this study a practical solution was proposed for determining the minimum amount of pathologist annotations leveraging low-accuracy panCK annotations, and subsampling the pathologist annotation dataset. Experimental study design and sample procurement requirement for generating a representative training and testing dataset that captured the vast variations of NSCLC on H&E images (such as subtype, scanner, center, study, operator, etc.) were also addressed. Transfer learning was used here to leverage a large amount of panCk-based tumor annotations for initial training, followed by fine-tuning using a small amount of high-accuracy pathologist annotations. Moreover, a multi-center, multi-protocol, multi-scanner dataset was generated for training the segmentation model. This approach has the potential to significantly reduce the burden of manual annotation by pathologists and improve the generalizability of the trained model. The effect of panCK pre-training on the final model performance and its impact on determining minimal pathologist annotations were studied. The panCK-based pre-training was compared to using pre-trained weights from foundation models and the impact of using different architecture in model backbone (feature extraction step) on model performance was investigated. Furthermore, the effects of different approaches in generating panCK annotations, as well as train- and test-time color normalization on model performance were investigated.

## Experimental setup and methods

The image analysis approach depends on the objectives of the clinical study. If the goal is to identify whether tumor is present in a WSI, a classification analysis would be sufficient. If the team is interested in localizing the tumor in the WSI, but the exact location of small structures (e.g. whether a cell is in or outside the tumor) is not required object detection analysis, which is a combination of classification and localizing structures would be sufficient (e.g. if distance of tumor nests from lymphocytes clusters is of interest). However, if the analysis is

interested in specifying which cells in WSI belong to tumor and perform more advanced investigation on these cells (e.g. how much of a specific marker is expressed in tumor cells and how much is outside tumor), image segmentation that performs classification for each pixel in the WSI needs to be performed and is the objective of the current work.

## NSCLC samples

Human NSCLC samples were procured from Indivumed, Capital Biosciences, Asterand, Discovery Life Science, Cureline, BioIVT, and Tristar under institutional review board (IRB) and ethics committee approvals of the respective vendors. All experiments were conducted in accordance with the national and international guidelines. Informed consent was obtained from all subjects by their respective vendors and all experiments were conducted in accordance with the obtained IRB approvals. The study was carried out in accordance with the guidelines and principles of the Helsinki Declaration. NSCLC samples from four different studies were included, where the H&E and immuno-fluorescence (panCK) staining and scanning were performed on 5 µm thick formalin-fixed paraffin-embedded (FFPE) sections. A total of 112 samples were procured (61 LUAD and 51 LUSC). For 11 samples, only one section was available and was used for H&E imaging (Studies 1&2); for 21 samples, two sections were available and were used for H&E and panCK imaging (Study 3), and for 80 samples four sections were available and were used for panCK and H&E imaging at three different centers (Study 4). The IF images were used to generate panCK-based tumor annotations, and the H&E images were used for generating manual tumor annotations by pathologists. Table 1 provides details of the samples included in this study.

## Immuno-fluorescence annotations

Tissue staining with tumor marker panCK could be performed using either immune-histochemistry (IHC) or immuno-fluorescence (IF) staining. IF staining was selected here as it is possible to perform H&E imaging after IF imaging on the same slide. This allows for perfect alignment between the H&E and its corresponding panCK-based annotations. If IHC staining was used, the H&E had to be performed on a serial section which would result in unaltered H&E images but alignment between tumor and its annotation would not be as accurate. Additionally, identifying tumor in IF stained images is more accurate as the only signal in these images is from panCK, whereas in IHC the marker signal is added to the nuclei and tissue structure signals and quantification is not as straight forward. Each approach (creating annotation on the same slide and on a serial slide) has its pros and cons and thus, using IF staining, we were able to investigate both approaches and assess their impact on model performance.

The NSCLC segmentation model relies on H&E images, and having unaltered H&E would be beneficial in panCK-based annotation generation step. However, considering H&E is a very stable stain while IF is less stable and that H&E is intrinsically immuno-fluorescent which cannot be removed by destining H&E (i.e. H&E staining cannot be performed prior to IF staining), in practice IF staining is always performed first, followed by epitope retrieval and H&E staining[40].

A singleplex IF assay for Pan Cytokeratin (Millipore Sigma, USA, AE1/AE3, 1 ug/mL) was performed on the FFPE human NSCLC samples. AE1 recognizes CK10, 14, 15, 16, and 19, while AE3 recognizes CK1, 2, 3, 4, 5, 6, 7, and 8. The assay was developed using the Opal technology workflow (Akoya Biosciences) on the Bond RX autostainer (Leica), as previously described[41]. 5um thick sections were obtained from each sample, mounted on charged glass slides (Statlab), baked at 60 °C for 60 min, and loaded onto the autostainer. Then, heat induced epitope retrieval was performed using the autostainer's built in "HEIR 20 min with ER2" protocol, followed by the singleplex IF assay for panCK. Slides were then removed from the autostainer, coverslipped with ProLong Gold mounting media (ThermoFisher), and scanned on the Vectra Polaris scanner (Akoya Biosciences) at 20x magnification.

The IF scan was followed by H&E staining and scanning of the same section at 40x magnification using a Leica Aperio ScanScope (AT2) scanner. The images were downsampled to 20x magnification (using bi-cubic interpolation) to be at the same resolution as the panCK images. The IF staining involved heat induced epitope retrieval followed by decoverslipping and re-staining with H&E. Thus, the H&E images acquired after IF scanning (hereby

| Study number | Processing center | Available tissue | LUAD | LUSC | H&E | IF | Scanner | Pathologist annotation |
|---|---|---|---|---|---|---|---|---|
| Study 1 | Center 1 | Section 1 | – | 5 | H&E-only | – | AT2 | 5 |
| Study 2 | Center 1 | Section 1 | – | 6 | H&E-only | – | AT2 | 6 |
| Study 3 | Center 1 | Section 1 | 21 | – | H&E-after-IF | panCK | AT2, Polaris | – |
| | | Section 2 | 21 | – | H&E-only | – | AT2 | 21 |
| Study 4 | Center 1 | Section 1 | 40 | 40 | H&E-after-IF | panCK | AT2, Polaris | – |
| | Center 1 | Section 2 | 40 | 40 | H&E-only | – | AT2 | 16 |
| | Center 2 | Section 3 | 40 | 40 | H&E-only | – | AT2 | 16 |
| | Center 3 | Section 4 | 40 | 40 | H&E-only | – | P250 | 16 |

**Table 1.** Details of the NSCLC samples including the number of samples for each subtype from each center and study, the image type (H&E or IF), scanner type, and the number of samples that were annotated by pathologists.

called H&E-after-IF) had slightly different color space and some morphological changes to the tissue occurs compared to a slide that had only been stained and scanned for H&E (hereby called H&E-only) as shown in Fig. 2.

The panCK images were thresholded using Visiopharm software (Hoersholm, Denmark) to generate tumor annotations and, along with the H&E-after-IF WSI, were used as the image and label pair for training. The 80 samples in Study 4 (40 LUSC and 40 LUAD), as well as the 21 samples in Study 3 (21 LUAD) were used to generate this dataset. After quality control of the panCK and H&E-after-IF images, 68 samples were selected, and panCK-based tumor labels were generated. A non-pathologist excluded normal lung from these images by drawing a rough boundary of the normal tissue. PanCK positive tissue labels were generated using adaptive thresholding of the panCK signal in Visiopharm software. This process involved detecting tissue areas using the DAPI, panCK and Autofluorescence channels. For adaptive thresholding of the panCK signal, the signal was first normalized over a $500 \times 500$ pixel area to manage signal gradients in panCK images. Then, any pixel that had normalized value greater than 1.4 was considered positive (this threshold was determined empirically). The panCK positive area was then smoothed to fill small holes (panCK is a membrane marker and cell nuclei have no signal). These masks were then exported from Visiopharm and used as the annotation for the H&E WSI.

It is worth noting that since in a typical clinical study panCK is usually not available, the input to the segmentation model is only an H&E image. PanCK in only used for generating tumor annotation and is only required in training the segmentation model, while during inference only an H&E image is required.

## Pathologist annotations

Histological slides were stained with H&E and scanned by either Leica Aperio ScanScope (AT2) or 3DHISTECH Panoramic 250 (P250) digital scanners at 40x magnification. This data was downsampled to 20x magnification (same as the IF annotation step) to generate the H&E-only images. There is significant variation in H&E images due to differences in staining and imaging protocols, scanners, centers, operators, and tumor subtypes. Incorporating all of these variations into the training data is challenging, which makes training a robust CNN model difficult (i.e., requires a large and diverse training dataset).

In order to incorporate variations in staining and scanning, sections 2–4 of the 80 samples in Study 4 were sent to three different centers (centers 1–3), where they used their respective scanners (AT2 and P250) and protocols (protocols 1–3) for H&E staining and imaging. Additionally, 11 LUSC (Studies 1–2) and 21 LUAD (Study 3) samples that were prepared and stained by different operators and scanned with different AT2 scanners were also included to increase H&E data variability. Figure 1c shows example images from each dataset, demonstrating the significant variation that can be expected in H&E images of NSCLC.

Out of these 112 images, 80 H&E images were selected for manual annotation by pathologists. The selected samples included 40 LUAD and 40 LUSC samples. Centers 1 & 2 used the same scanner type, and thus there were 58 samples scanned with AT2 scanners, and 22 samples were scanned on the P250 scanner. For each of the 80 H&E WSI, three 1 $mm^2$ regions of interest (ROI) were selected for manual annotation by pathologists using QuPath software[42]. This selection that included both LUAD (50%) and LUSC (50%) samples was confirmed by an expert pathologist to ensure a wide range of NSCLC tissue morphologies, normal, and tumor-adjacent tissues were included in the training data. The pathologists drew contours around the tumor cells at high magnification (20x) and excluded any non-tumor cells and structures such as stroma, normal lung cells, normal epithelium, blood vessels, lymphocytes, etc. To incorporate variability in the annotation between pathologists (due to years

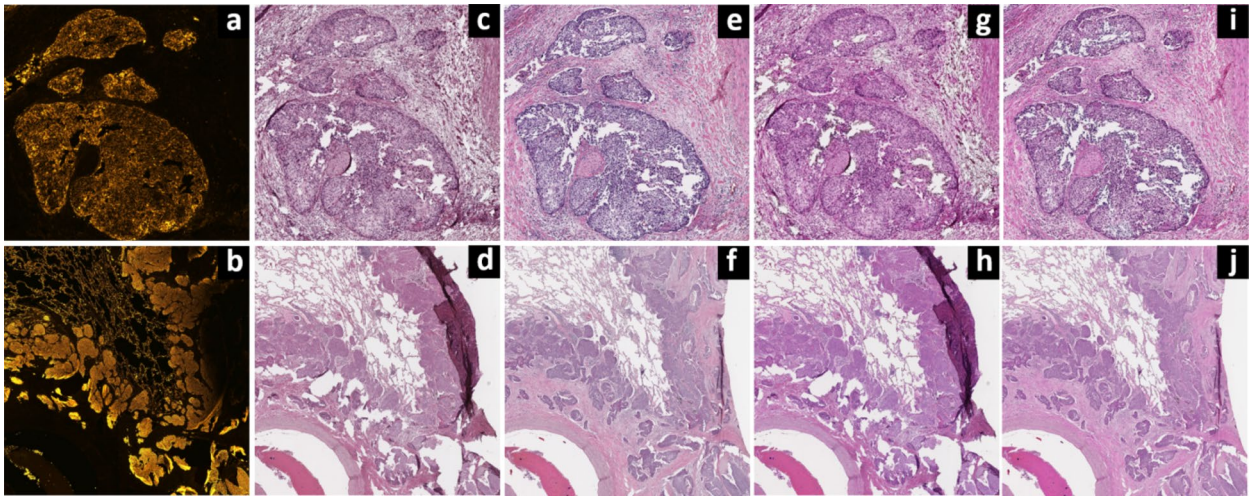


**Figure 2.** Sample images of panCK (**a**, **b**), H&E-after-IF (**c**, **d**), and corresponding H&E-only image from a serial section (**e**, **f**) showing the differences in color space between H&E-after-IF and H&E-only images. The images also show the morphological changes in the tissue due to IF staining (particularly in non-tumor tissues) and re-coverslipping (tissue fold and tear), as well as co-registration issues (mismatch between H&E-only and panCK). Color normalization transforms these images to a similar color space: H&E-after-IF (**g**, **h**), H&E-only (**i**, **j**).

of experience and differences in their judgment calls), the images were annotated at two different centers, and at each center, multiple pathologists annotated the slides.

Out of the 80 annotated H&E WSI samples, 48 samples were selected from Study 4 samples (16 from each center), and 32 WSI were selected from Studies 1–3 (all performed at Center 1). CNN model training was performed using 58 WSI images (29 LUAD & 29 LUSC), and the remaining 22 WSI (11 LUAD & 11 LUSC) were kept for testing (covering all subtypes, centers, protocols, and scanners).

The pathologists annotated 174 ROIs of size 1 $mm^2$ in training/validation data. In addition, 72 non-tumor ROIs of size 1 $mm^2$ from these WSI were also included in the training data. Out of the total 246 ROIs, 15% were used as the validation dataset (36 ROIs), and the remaining 210 ROIs were used as training dataset (covering 210 $mm^2$ over 58 WSI). This dataset was downsampled randomly (at ROI level) at 50% (105 ROIs), 30% (70 ROIs), 20% (42 ROIs), and 10% (21 ROIs), and for each fraction, five random subsets were created resulting in 21 training sets (including the one dataset with 100% of the training data).

The test dataset was comprised of 66 ROIs of size 1 $mm^2$ from 22 WSI, which were annotated by the pathologists. These ROIs were evenly distributed between the two subtypes (33 LUAD and 33 LUSC) and the three centers.

## CNN architecture

The CNN used for NSCLC tumor segmentation was based on the attention U-Net[43,44] which is a modified version of the U-Net architecture[45] with attention gates added to each resolution level of its decoder. Using models with pre-trained weights have been shown to outperform those trained from scratch in handling out of distribution images in digital pathology[46]. Thus, pre-trained weights of VGG16 network trained on ImageNet dataset were used for the encoder half of the attention U-Net architecture (Fig. 3).

Inputs to the model were H&E patches of size $512 \times 512 \times 3$ at 20x magnification. The original H&E images that were acquired at 40x were downsampled using bi-cubic interpolation to arrive at the 20x H&E images and a $5 \times 5$ convolutional kernel was used in convolutional layers of the model decoder. The VGG16 architecture that

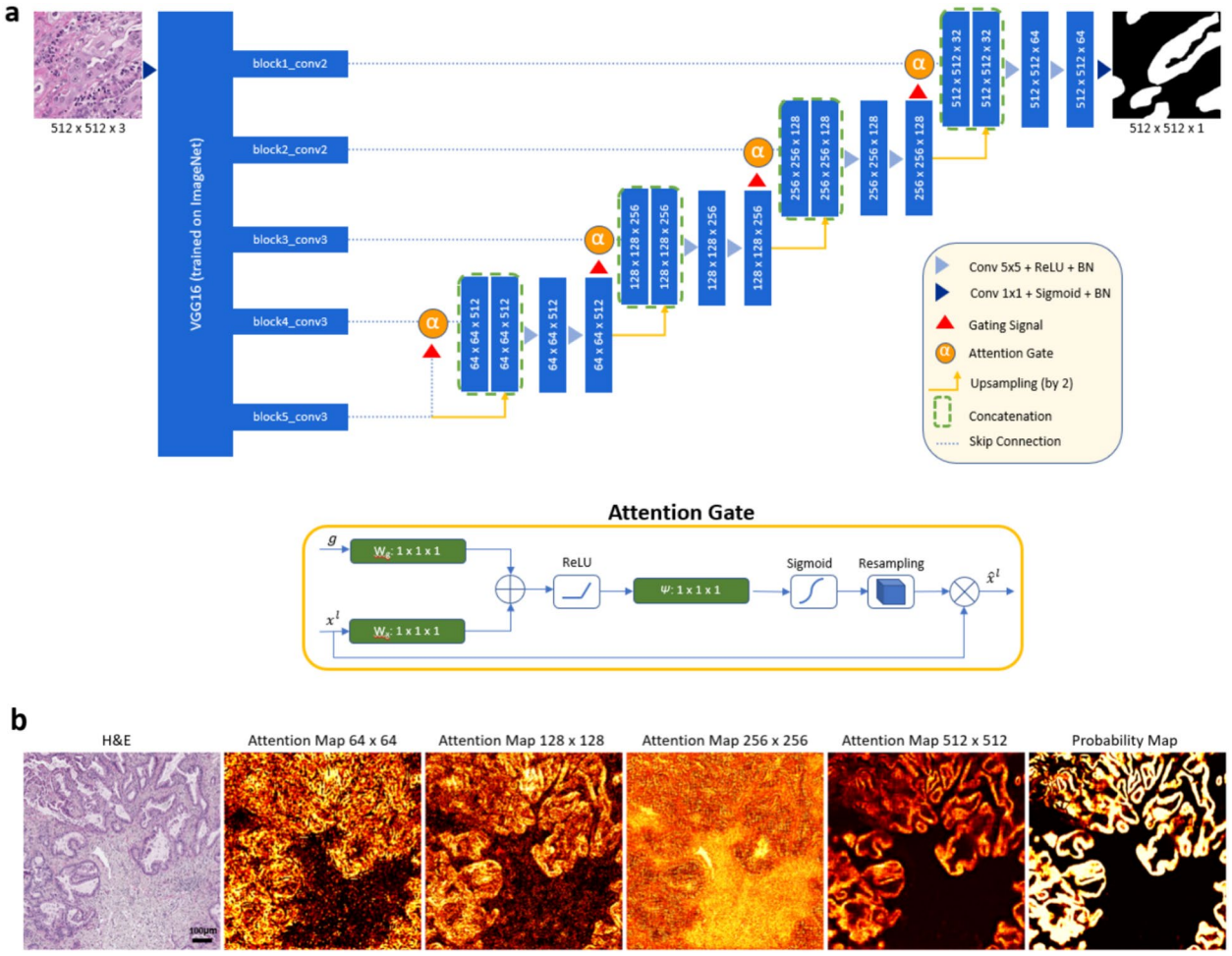


**Figure 3.** The attention U-Net architecture with pre-trained weights of VGG16 on ImageNet dataset (**a**). For a sample NSCLC image, the output of attention gates at different levels of the network are shown. These maps show how the attention gates are placing emphasis on the tumor tissue while minimizing the weighting for background tissues (**b**).

is used in the encoder has 3 × 3 kernel in its convolutional layers which results in a receptive field of 212 × 212[47]. However, 512 × 512 images were used as input here and thus, the kernel size of the decoder was increased to increase the receptive field of the full network. The rationale for selecting this larger input size (which is commonly used in the literature for WSI[46]) was the fact that tumor segmentation requires some global context to differentiate structures that are similar at cell level (e.g. tumor cells that are epithelial and normal epithelium).

Model details are shown in Fig. 3. Regular U-Net with the same model architecture and parameters as the attention U-Net (without attention gates) was also implemented and used for comparison.

## CNN training and transfer learning

The model was trained in 2 steps where initially, the weights in the encoder half of the model were frozen, and only the decoder was trained with initial learning rate of LR = 0.001 and LR was reduced in half if the validation loss did not decrease for 4 epochs. Training was stopped if the validation loss did not decrease for 15 epochs and the model with the lowest validation loss was selected. In the second step, the encoder was also trained with an initial learning rate of LR = 0.0001 and the same LR reduction schedule and early stopping criteria. Binary cross entropy (BCE) loss was used and the tensorflow package (http://www.tensorflow.org) was used for implementing and training the CNN. The model was first trained using the panCK-based tumor annotations and H&E-after-IF images. This model was then fine-tuned using the pathologist annotations on H&E-only images.

To increase model generalizability, image augmentation including rotation (up to 45°), horizontal and vertical shift (up to 20%), zooming into the image (up to 5%), shear deformation (up to 10%), horizontal and vertical flips, as well as random 90°, 180°, 270° rotations were applied to the image patches during training. WSI images have no natural orientation and using the combination of 45° rotation along with random 90°, 180°, 270° rotations and horizontal and vertical flips ensured the image patches were rotates for the all possible rotation degrees. Additionally, to increase variability in color space, color augmentation was performed using the stain intensity (Hematoxylin and Eosin) perturbations technique presented by Tellez et al.[48,49], which is based on the Macenko color normalization approach[50].

## Architecture backbone

In order to assess the impact of backbone architecture (which represents the feature extraction phase of the model) on the model performance, several pre-trained backbones were tested. Resnet50[27], DenseNet121[28], and EfficientNet-B4[29] pre-trained on ImageNet dataset were used as backbone in addition to VGG16. Moreover, we trained Swin-UNETR architecture[30] to compare the performance of the architectures that used CNN vs. transformers (Swin Transformer here[51]) for feature extraction. The MONAI platform[52] was used for Swin-UNETR model implementation. All models were trained similarly where first the model was trained using the panCK annotations followed by fine-tuning using the pathologist annotations.

## Pre-training: panCK vs. foundation models

The current training approach used panCK for pre-training the model. Foundation models are another approach for pretraining. Thus, the performance of using foundation models vs. the task and domain specific pre-training using panCK annotations was performed. Two foundation models, KimiaNet[24] which is a CNN based model (DenseNet121), and CTransPath[25] which is transformers-based model were used. Both foundation models were trained using pathology images, and for both cases, the model was trained for a classification task. Thus, the models required being adapted for our segmentation task where only the backbone (feature extraction) part of the segmentation models was pre-trained in the foundation model and the decoder part of the model required training.

For KimiaNet that used DenseNet121, the same approach as the case of using DeneNet121 pre-trained using ImageNet data was used and only the KimiaNet weights were used as the backbone. Thus, the same attention U-Net architecture was used. For the case of CTransPath that used Swin transformers, the model was adapted for segmentation using the mask2former approach[53]. These two models were trained using the pathologist annotations only.

## Evaluation

Three pixel-based model performance metrics were calculated. (1) accuracy which calculates the ratio of true positives (tumor pixels correctly classified as tumor) and true negatives (background pixels correctly classified as background) over all pixels in the image. Considering majority of pixels in WSI are usually background, accuracy places higher weight on non-tumor regions and if a small tumor region is missed, the metric does not properly penalize it; (2) mean intersection over union (mIoU) which represents the ratio of the true positive pixels, divided by the sum of true positives, false positives (background pixels incorrectly classifies as tumor) and false negatives (tumor pixels incorrectly classified as background). This metric is preferred for image segmentation tasks as it properly penalize the metrics if only a small tumor region exists in the image; (3) Dice coefficient which is similar to mIoU but with a different formulation. We are reporting both Dice and mIoU in this paper to facilitate comparison with prior studies (some studies in the literature report mIoU and others report Dice which makes comparing different studies problematic). All three metrics were calculated for each 1 $mm^2$ ROI in the test dataset that was comprised of 66 ROIs from 22 WSI. Overall performance was determined as mean and standard deviation or median and interquartile range on the entire test dataset, as well as separately for different tumor subtypes and the centers.

### Minimum required pathologist annotations

Generating pathologist annotations is the main bottleneck in developing segmentation models in histopathology due to being expensive and time-consuming. To evaluate the effects of pre-training the CNN with panCK-based tumor annotations, training was performed with and without panCK-based pre-training.

Moreover, to determine the minimum amount of pathologist annotations that would be required to train a model with acceptable accuracy, the pathologist annotation dataset was subsampled at 10%, 20%, 30%, and 50%, and the model was trained using these smaller datasets. To preserve the heterogeneity in these smaller datasets, sub-sampling was performed at ROI level (the training data included 246 ROIs of size 1 $mm^2$ annotated on the 58 training WSI). For each sub-sampling fraction, five random datasets were generated. For the cases that 100% of the training data was used, the train/validation split was performed randomly five times. The model was trained with and without using the panCK-based pretraining. The same approach was used for models using pre-trained backbone from foundation model weights (KimiaNet and CTransPath) to compare the effects of using panCK vs. foundation model weights for pre-training while reducing the size of the pathologist annotations dataset. The performance of each model was then tested on the same independent test dataset (66 ROIs from 22 WSI).

## Results

The objective was to (a) investigate different approaches for generating ground truth annotations for NSCLC tumor segmentation on H&E and determine best practices for such data procurement; and (b) determine the minimum required pathologist annotations by leveraging panCK annotations which is easy to generate at scale. Thus, a comprehensive set of NSCLC slides were procured containing both LUAD and LUSC subtypes and were H&E stained and scanned at three centers using different protocols, scanners, and operators. This dataset was used to initially generate a large dataset of pathologist annotations to provide segmentation accuracy in-line with the literature. Then, panCK annotations were used along with a subset of the pathologist annotations to determine the minimum required size of pathologist annotations. These datasets were generated while considering the factors that impact model generalizability and the challenges each factor creates (e.g., differences in H&E-after-IF quality compared to H&E-only, challenges associated with panCK annotations, etc.). Moreover, the current approach, that used panCK annotations for model pre-training, was compared to using foundation models for pre-training the model backbone.

### Training using panCK annotations

The attention U-Net model (using VGG16 as backbone) was trained using panCK-based tumor annotations and H&E-after-IF images. The total panCK annotated tissue area used for training was 10,326 [$mm^2$]. The model trained on this data resulted in a mean intersection over union (mIoU) of $67 \pm 11$ [%] and a dice coefficient of $80 \pm 8$ [%] on the test dataset. Figure 4 shows the results of applying the model to a LUSC and a LUAD sample.

The epitope retrieval process in panCK staining involves heating the slide to 95 °C degrees for approximately 20 min. This process causes changes in the color space and tissue morphology on H&E (particularly in the non-tumor tissues). Tissue loss, tear, and fold can also occur during the coverslip removal and re-coverslipping of the slides for H&E staining after IF staining and imaging. Figure 2 shows an example of an NSCLC slide stained with panCK (Fig. 2a,b) followed by epitope retrieval and re-staining with H&E (H&E-after-IF, Fig. 2c,d) as well as the H&E-only image of its serial section (Fig. 2e,f), showing the extent of the changes in H&E due to this process.

The change in color space can be reduced by color normalization (Fig. 2g,h), which transforms the images to a color space similar to color-normalized H&E-only images (Fig. 2i,j). However, there exist some structural/morphological changes in the non-tumor tissues. The tumor morphology on H&E-after-IF images is less impacted and has a similar imaging appearance to the H&E-only images. These issues, as well as the non-specificness of panCK labels to tumor tissue (Fig. 1a) and technical challenges in panCK imaging (Fig. 1b), impacted the performance of the model and resulted in sub-optimal segmentation results.

### Fine-tuning using pathologist annotation

The model that was trained on panCK labels was fine-tuned using the pathologist annotations. The pathologist annotation dataset (246 [$mm^2$]) was significantly smaller than the panCK labels (10,326 [$mm^2$]). Thus, initially the encoder was frozen and only the decoder weights were updated. In the second step the encoder wights were also made trainable and the entire model weights were updated with a small learning rate. Figure 3b shows the attention maps at different resolution levels of the final fine-tuned model for a LUAD sample. Figure 4 shows improved segmentation results after fine-tuning for a LUAD and a LUSC sample. Additionally, Fig. 5 shows the segmentation results for images from different scanners, studies, protocols, and tumor subtype showing the improved segmentation performance of the model. The final model resulted in mIoU of $81 \pm 10$ [%] and Dice coefficient of $89 \pm 7$ [%] on the entire test dataset with no test-time color normalization. The same performance metrics were achieved when test-time normalization was used (mIoU of $81 \pm 10$ [%] and a Dice coefficient of $89 \pm 7$ [%]). Thus, color-normalization during training is sufficient.

The model used VGG16 weights trained on ImageNet dataset as backbone and to assess the impact of the backbone architecture on model performance, several architectures were also used. Table 2 reports the model performance for using these architectures in the model backbone (panCK pre-training followed by fine-tuning using the entire pathologist annotations dataset). Considering the similar performance of the model for different backbones, VGG16 was used as the backbone in this study. Foundation models were also used as pre-trained backbone and the model performance when the entire pathologist annotation dataset was used (without panCK pre-training) is reported in Table 2.

To assess the impact of having attention gate in the model architecture, regular U-Net with the same model architecture and parameters as the attention U-Net model (without attention gates) was also trained. The regular

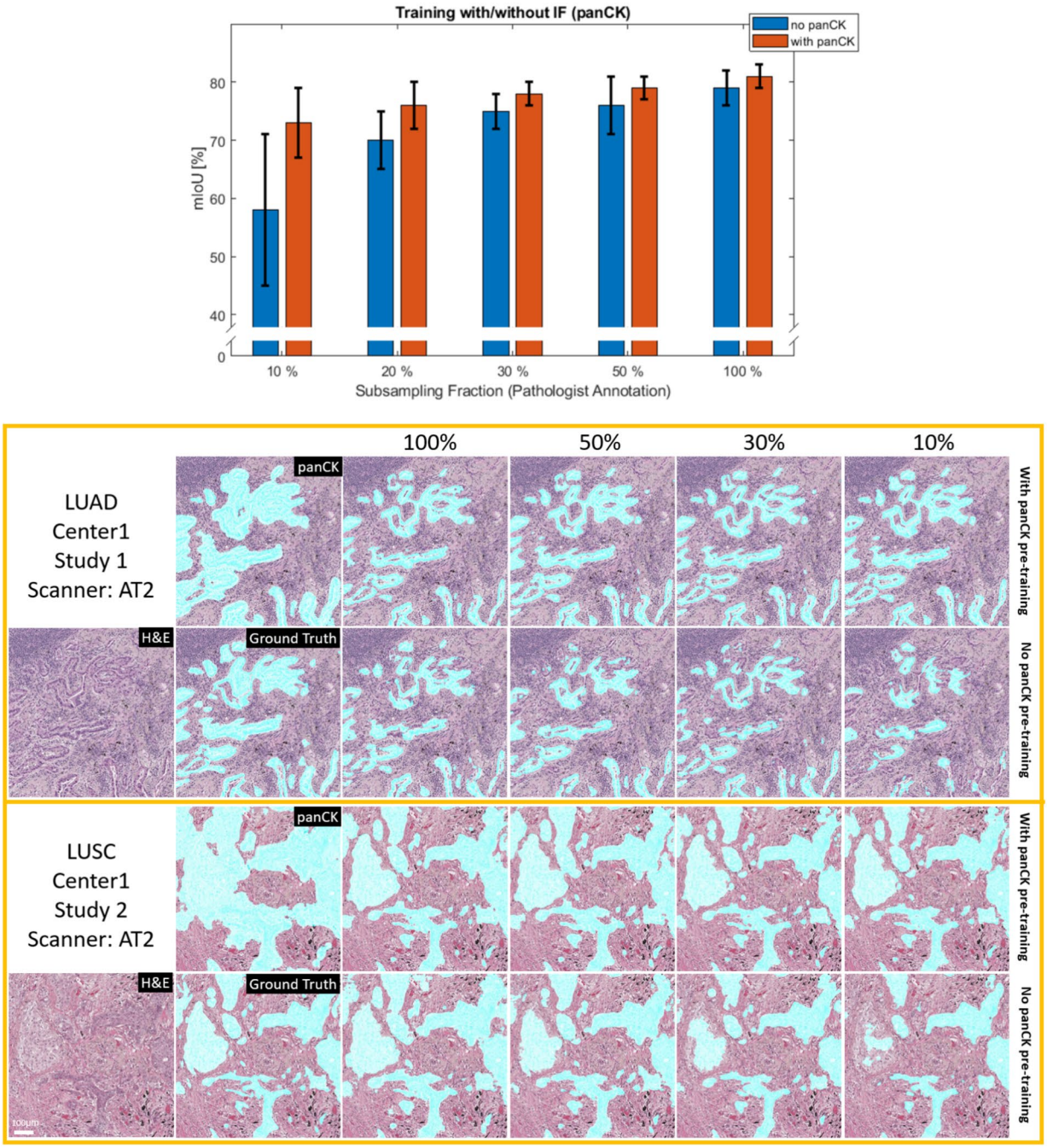


**Figure 4.** Barplot shows the model performance (mIoU) change as the pathologist annotation dataset size is reduced from 100% to 10% with and without pre-training with panCK annotations. Model performance on a LUAD and a LUSC sample is shown for the models with different sizes of pathologist annotations, with and without pre-training using panCK annotations. The ground truth annotations by pathologists as well as the performance of the panCK pre-trained model (without pathologist annotations) are also shown.

U-Net achieved mIoU of 79 ± 11 [%] and Dice coefficient of 88 ± 7 [%], which is very close to the performance of the attention U-Net model and considering the standard deviations of these performance metrics, suggests no significant improvement is achieved by adding the attention gates.

Table 3 reports the final model performance segregated for different sub-types and the centers that performed the staining and scanning, showing the model performed similarly across different subtypes, scanners, centers, and protocols.

## Minimum pathologist annotation requirement

A total of 21 training datasets were prepared, and for each dataset the model training was performed with and without using the panCK pre-training, while keeping everything else (model, hyper-parameters, preprocessing)

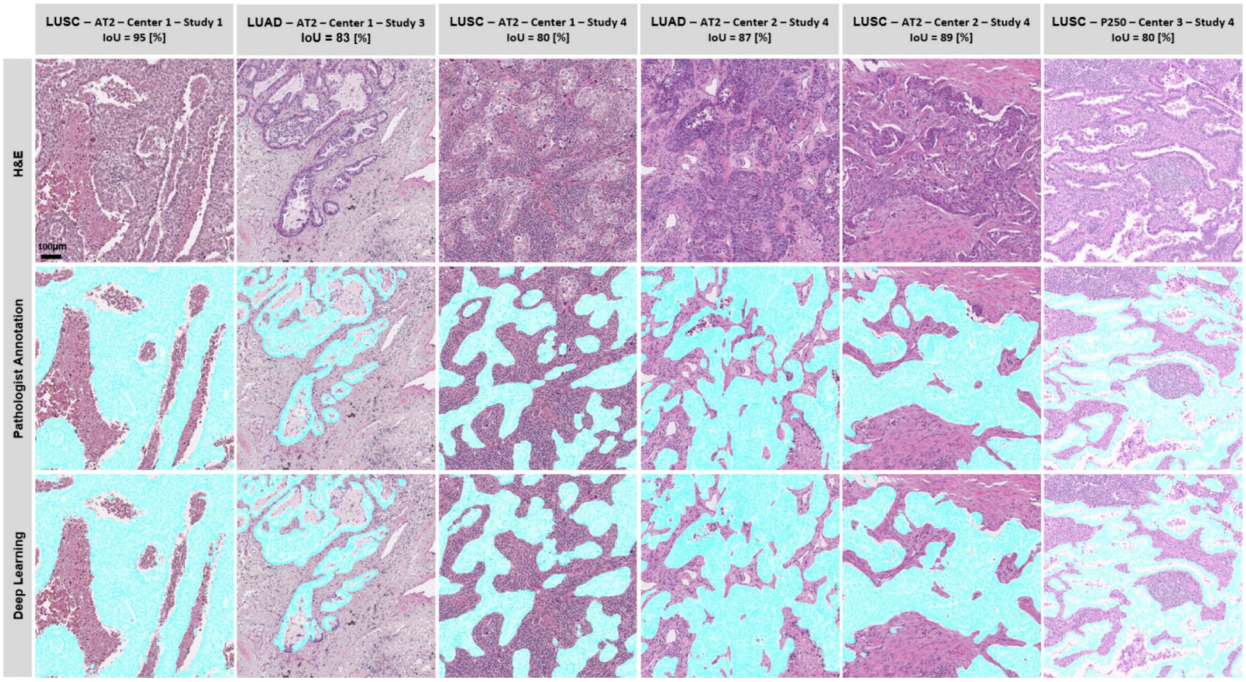


**Figure 5.** Segmentation results of the model pre-trained with panCK annotations followed by fine-tuning with 100% of the pathologist annotations on representative images from both subtypes (LUAD and LUSC), both scanner types (AT2 and P250), all three centers (different protocols and scanners), and multiple studies (different operators), showing the model is robustly segmenting the tumor in such a diverse dataset (diversity in morphology and color space).

| Architecture | Backbone | PreTraining | mIoU | Dice | Accuracy |
|---|---|---|---|---|---|
| Attention U-Net | VGG16 | panCK | 81 ± 10 | 89 ± 7 | 92 ± 3 |
| Attention U-Net | DenseNet121 | panCK | 82 ± 9 | 90 ± 5 | 92 ± 2 |
| Attention U-Net | Resnet50 | panCK | 81 ± 9 | 89 ± 6 | 92 ± 4 |
| Attention U-Net | EfficientNet-B4 | panCK | 82 ± 9 | 90 ± 6 | 92 ± 3 |
| Swin-UNETR | Swin transformer | panCK | 78 ± 10 | 87 ± 6 | 89 ± 4 |
| U-Net | VGG16 | panCK | 79 ± 11 | 88 ± 7 | 91 ± 4 |
| Attention U-Net | KimiaNet (DenseNet121) | Foundation model (backbone only) | 79 ± 12 | 88 ± 9 | 91 ± 5 |
| Mask2Former | CTransPath (Swin transformer) | Foundation model (backbone only) | 77 ± 8 | 87 ± 6 | 89 ± 5 |

**Table 2.** Model performance using different backbone architectures (when using the entire pathologist annotations dataset). Pre-trained weights of backbones on ImageNet dataset was used and models were first trained using panCK annotations followed by fine-tuning on pathologist annotations. For foundation models, pre-trained weight from the foundation model was used as backbone and only the pathologist annotations dataset was used for training (i.e. no training on panCK annotations).

| | ALL | Adenocarcinoma (LUAD) | | | | Squamous cell carcinoma (LUSC) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ALL | Center 1 | Center 2 | Center 3 | ALL | Center 1 | Center 2 | Center 3 |
| mIoU | 82 [77 87] | 82 [75 87] | 82 [70 85] | 84 [81 90] | 77 [70 86] | 83 [79 87] | 84 [73 88] | 83 [82 87] | 81 [79 87] |
| Dice | 90 [87 93] | 90 [85 93] | 90 [83 92] | 91 [89 95] | 87 [83 93] | 90 [88 93] | 91 [84 93] | 91 [90 93] | 90 [88 93] |
| Accuracy | 92 [90 94] | 91 [89 93] | 92 [91 94] | 91 [89 94] | 90 [88 92] | 92 [90 94] | 93 [91 94] | 90 [89 93] | 92 [90 95] |

**Table 3.** Median and inter-quartile range for performance metrics of the model calculated for different subsets of the test dataset.

identical. For the case of no panCK pre-training, the encoder layers (pre-trained wights of VGG16 on ImageNet) were fixed first, and only the decoder was trained. Once the model converged, the encoder layers were also made trainable and the entire model was fine-tuned using a smaller learning rate (similar process to training the model using panCK annotations only). Similar approach was taken for using foundation model wights as the backbone. The model performance is reported in Table 4.

When using the entire training dataset, the model pre-training using panCK resulted in mIoU of 81 ± 10 [%] and Dice coefficient of 89 ± 7 [%], compared to mIoU of 79 ± 13 [%] and Dice coefficient of 87 ± 10 without panCK pre-training. For regular U-Net (no attention gates) mIoU of 79 ± 11 [%] and Dice coefficient of 88 ± 7 [%] was achieved with panCK pre-training, and mIoU of 77 ± 15 [%] and Dice coefficient of 86 ± 12 [%] was achieved without panCK pre-training. Thus, the best performance was achieved when using Attention U-Net with panCK pre-training. As reported in Table 4 and shown in Fig. 4a, the model performance dropped sharply as the training data size was decreased when no panCK pre-training was involved, however, the model performance was stable and had a much smaller drop when panCK pretraining was used. Using foundation model weights (which only provided pre-trained weights for the model backbone and the decoder required training from scratch), showed similar drop in model performance (similar to no panCK pre-training) compared to pre-training the entire model using panCK annotations. Figure 4b shows how the model performance dropped as the pathologist annotation dataset size was reduced (with and without panCK pre-training) for a LUSC and a LUAD sample. This figure shows using panCK pre-training resulted in more robust tumor segmentation as the pathologist annotation dataset size was reduced.

## Discussions

Tumor segmentation on H&E eliminates the need for additional tumor markers (IHC or IF) in histological image analysis of the tumor tissue. CNNs use the morphological differences between tumor and non-tumor cells, present in H&E WSI, for tumor segmentation. However, the main bottleneck in training a CNN lies in generating sufficient high-quality ground truth tumor annotations. Pathologist annotations are very expensive and time consuming, limiting their feasibility to small datasets. Alternatively, IF or IHC markers can be used to generate tumor annotations, and extensive efforts have been made to employ panCK, the most used tumor marker in histopathology, for annotating large datasets at a relatively low cost. Nevertheless, panCK's accuracy is insufficient for training a robust CNN model for NSCLC tumor segmentation[31–33]. An alternative to panCK-based pre-training is using the pre-trained weights of foundation models that have been trained on a very large dataset of histological images[24,25,54].

This study explored the potential of using the low cost panCK-based annotation on a large dataset for pre-training the model, followed by fine-tuning using a small dataset of pathologist annotations, and compared it to using foundation models for model pre-training. The shortcomings of panCK-based annotations were identified and their impact on training a segmentation model was assessed. The impact of panCK on determining the minimum size of pathologist annotation dataset was studied. The impact of using various architectures as model backbone for feature extraction was also explored. To train a generalizable model, the sources of variability in H&E images which include differences in scanners, staining protocols, centers, and operators, as well as the NSCLC subtypes were also considered. The current study provided a comprehensive evaluation on the study design considerations and provided a practical solution for conducting such an experiment successfully.

### Model architecture

Various choices for model backbone were examined. Table 2 reported the model performance when using VGG16, ResNet50, DenseNet121 and EfifcientNet-B4 as the attention U-Net backbone, as well as Swin Transformer through the Swin-UNETR architecture. This table showed there were negligible differences between model performances arising from the backbone architecture (except for Swin-UNETR that had lower performance

| Training data percentage | Pre-training | panCK | 100% 210 [mm$^2$] | 50% 105 [mm$^2$] | 30% 70 [mm$^2$] | 20% 42 [mm$^2$] | 10% 21 [mm$^2$] |
|---|---|---|---|---|---|---|---|
| IoU [%] | with panCK | 67 | 81 ± 2 | 79 ± 2 | 78 ± 2 | 76 ± 4 | 73 ± 6 |
| | no panCK | N/A | 79 ± 3 | 76 ± 5 | 75 ± 3 | 70 ± 5 | 58 ± 13 |
| | KimiaNet | N/A | 79 ± 2 | 76 ± 4 | 75 ± 2 | 73 ± 3 | 65 ± 10 |
| | CTransPath | N/A | 77 ± 4 | 73 ± 2 | 72 ± 3 | 68 ± 5 | 61 ± 9 |
| Dice [%] | with panCK | 80 | 89 ± 3 | 88 ± 1 | 87 ± 1 | 86 ± 3 | 83 ± 5 |
| | no panCK | N/A | 87 ± 2 | 85 ± 4 | 85 ± 2 | 81 ± 5 | 69 ± 13 |
| | KimiaNet | N/A | 88 ± 3 | 85 ± 3 | 85 ± 2 | 83 ± 3 | 75 ± 11 |
| | CTransPath | N/A | 87 ± 3 | 82 ± 3 | 81 ± 2 | 77 ± 5 | 70 ± 10 |
| Accuracy [%] | with panCK | 80 | 92 ± 2 | 91 ± 1 | 91 ± 1 | 90 ± 2 | 88 ± 3 |
| | No panCK | N/A | 91 ± 2 | 89 ± 2 | 89 ± 2 | 87 ± 3 | 81 ± 6 |
| | KimiaNet | N/A | 91 ± 2 | 89 ± 1 | 89 ± 1 | 88 ± 2 | 83 ± 7 |
| | CTransPath | N/A | 89 ± 3 | 86 ± 2 | 86 ± 2 | 83 ± 3 | 80 ± 2 |

**Table 4.** Mean and standard deviation of the model performances metrics for varying sizes of pathologist annotations dataset.

metric). Thus, VGG16 was selected as the backbone of the attention U-Net architecture in the remainder of this study.

Attention U-Net architecture, which incorporates attention gates into the skip connections of the U-Net was use here. These attention gates allow the model to place a greater emphasis on the tissue of interest, which in this case is the tumor, while reducing the weights given to the background, and potentially achieve more accurate segmentation results. Figure 3 provided an illustration of the output produced by the attention gates at all four resolutions for a LUAD sample. These attention maps demonstrated that the model effectively detected the tumor tissue at multiple levels and placed greater emphasis on the tumor area. Regular U-Net with the same architecture and parameters as the Attention U-Net was also used to assess the impact of attention gates on model performance showing attention gates lead to slightly higher performance metric which was not significant when considering the standard deviations of the two model performances (mIoU of 79 ± 11 [%] and mIoU of 81 ± 10 [%] for regular U-Net and Attention U-Net respectively).

## panCK-based model

Using panCk is an efficient approach for ground truth generation. However, it suffers from non-specificity of the panCK to lung tumor tissue only (normal lung, normal epithelium, necrotic tissue, and red blood cells also express panCk) as shown in Fig. 1a. In addition, there exist technical challenges in panCK imaging such as image gradient and signal drop (Fig. 1b) that may lead to inaccurate tumor labels, particularly if a threshold is being used in panCK images to separate the background from the tumor signal.

PanCK imaging can be conducted on either a tissue section serial to the H&E or on the same section. When utilizing a serial section, there are challenges related to the imperfect alignment of tumor labels with the H&E, particularly if the tumor is scattered within the tissue. Thus, co-registration of the IF and H&E images becomes necessary. Conversely, using the same section for both panCK and H&E staining involves a staining process where slides are first stained with panCK, followed by epitope retrieval and subsequent re-staining with H&E. This procedure can cause some damage to the tissue morphology, particularly in normal and non-tumor tissues, and also leads to a change in color space (Fig. 2). Nevertheless, this approach ensures a perfect alignment between tumor annotations and the H&E image, eliminating the need for co-registration. These issues reduce the accuracy of a model that is trained on panCK labels only.

The panCK-based tumor annotations were generated using both approaches, i.e., using the same section and a serial section. The CNN model was trained on both datasets, resulting in mIoU of 67 ± 10 and 66 ± 13, and Dice coefficient of 80 ± 8 [%] and 78 ± 10 [%] for the same section and a serial section, respectively. These performance metrics are similar to the performance metrics for using a serial section reported in the literature[31–33].

Fine-tuning these panCK-based models using pathologist annotations was required to improve results. After fine-tuning, the mIoU increased to 81 ± 10 [%] and 80 ± 10 [%], while Dice coefficient reached 89 ± 7 [%] and 89 ± 7 [%] for the same section and a serial section, respectively. Therefore, although there was slightly better performance when using the same section to generate panCK annotations, there was no significant difference in the model performance after fine-tuning using pathologist annotations.

These findings suggest that if only panCK annotations are available for training, it may be advantageous to use the H&E on the same section. This eliminated the need for an additional tissue section for H&E and also results in slightly improved model performance. Moreover, the process of generating tumor annotations becomes simpler as there is no requirement for co-registration.

## Pathologist annotation-based model

In Fig. 4, it is evident that the model trained solely on panCK annotations successfully detected tumors to a certain extent. However, it missed some parts of the tumor and incorrectly identified portions of the background and necrotic tissue as tumor. This observation emphasized the necessity of fine-tuning the model using pathologist annotations. After fine-tuning, the segmentation performance improved significantly, as illustrated in Fig. 4. When applied to the entire test dataset, the model mIoU increased to 81 ± 10 [%], and the Dice coefficient reached 89 ± 7 [%]. These performance metrics align with those reported in the literature[17,18], indicating the effectiveness of the proposed approach despite using a small dataset of pathologist annotations and the significant variability in the H&E image generation processes.

One of the major challenges in training a CNN on H&E images is the significant variation in color space caused by different stain concentrations and imaging parameters (e.g. exposure time) due to the use of different protocols (Fig. 1c). To overcome this issue, various approaches have been employed in the literature. Some studies have used test-time color normalization[19], while others have opted for color augmentation during training[23,36]. Test-time color normalization, such as the Macenko technique[50], involves color deconvolution and determining the color space eigenvectors, which can be time-consuming. Furthermore, it needs to be performed on every image during inference. This can quickly become burdensome if the model is to be used regularly on a large number of images and studies. In contrast, color augmentation during training is a one-time process and does not introduce any additional time during inference.

Therefore, we employed color augmentation during training based on the color normalization technique proposed by Macenko[50]. This approach perturbs the Hematoxylin and Eosin stain concentrations to generate new images[48,49]. We evaluated the model performance with and without test-time color normalization, and both scenarios yielded the same results: an mIoU of 81 ± 10 [%] and a Dice coefficient of 89 ± 7 [%]. This suggests that there is no additional benefit to using test-time color normalization when color augmentation is performed during training.

The test dataset used in this study consisted of H&E images obtained from three different centers, using multiple protocols and scanners. The dataset was evenly divided between LUAD and LUSC, enabling the evaluation

of the model performance on these distinct subsets, thus assessing its generalizability. The results presented in Table 3 indicate that the model performance remained consistent across all of these subsets, and Fig. 5 shows several examples of the segmented tumor across various subsets. These results demonstrate the model's robustness to variations arising from changes in tumor subtypes, H&E staining and scanning centers, scanners, protocols, and studies.

## panCK pre-training impact on minimal pathologist annotation

The results presented in Table 4 and shown in Fig. 4 demonstrate the positive impact of panCK pre-training on model performance, resulting in a 3-percentage-point improvement compared to the model without pre-training (when using 100% of the pathologist annotations). Additionally, the standard deviation of the performance metrics was reduced, indicating enhanced consistency. This improvement was particularly pronounced when the size of the pathologist annotation dataset was reduced.

Comparing the attention U-Net model training with and without panCK pretraining, when panCK pre-training was employed, the model's mIoU decreased from 81 to 77% as the pathologist annotation dataset size decreased from 100% to 10%. In contrast, when panCK pre-training was not used, the model's mIoU dropped from 78 to 65% under the same reduction in dataset size. Figure 4 also showcases the significant decline in segmentation accuracy when panCK pre-training was absent.

Table 4 also reported the model performance when using the pre-trained weights from two foundation models as the backbone: a CNN-based model (KimiaNet), and a transformers-based model (CTransPath). Using KimiaNet as the backbone showed some improvement in model performance compared to using pre-trained weights on ImageNet dataset, where the model resulted in smaller variations in performance metrics and higher metric values, particularly when the pathologist annotation dataset was very small (10% or 20% of the entire dataset). However, the performance metrics were significantly lower than the cases that used panCK annotations for pre-training. This table also showed that transformer-based pre-training in foundation model performed poorly compared to CNN-based models. This lower performance was also observed in Table 2 where the lowest performance metric was achieved for Swin-UNETR compared to CNN-based backbones. The lower performance of the foundation models when used as backbone of the segmentation model could be attributed to the fact that they only covered the model backbone and a significantly large part of the model still required training from scratch. However, when pre-training using panCK annotation, the entire model was pre-trained (not just the backbone), which in combination with panCK being specific to the segmentation task being addressed, resulted in better model performance and stability when reducing the size of the pathologist annotations dataset.

Thus, panCK pre-training not only enhanced model performance but also resulted in a more robust model, even in scenarios where there was limited availability of pathologist annotations. It is important to note that all models started with pre-trained weights of VGG16 network on the ImageNet dataset as the CNN encoder. Similar poor performance trends were also observed when using foundation model that were pre-trained on a very large dataset of histological image. These results demonstrate the domain-specific pre-training using the large dataset of panCK-based annotations played a crucial role in improving model performance.

Moreover, the results highlight that when using panCK pre-training, a smaller dataset of pathologist annotations could be used with minimal impact on model performance. Taking into consideration that three 1 $mm^2$ ROIs were annotated for each H&E WSI, a reasonable cutoff point would be using 30% of the pathologist annotations (mIoU of $78 \pm 2$ and Dice coefficient of $87 \pm 1$), equating to having one 1 $mm^2$ ROI annotated by the pathologist for each H&E WSI. This approach preserves the heterogeneity of the training data (using as many WSIs as possible) while minimizing the resources required for pathologist annotations.

## Limitations and future directions

A major limitation of the study was the size of the pathologist annotation dataset. Although we attempted to cover as much heterogeneity as possible in the training and test datasets, the variability and factors that impact H&E image quality and appearance are vast and diverse. NSCLC is a very heterogeneous tumor type with significant variations in its morphology. Capturing as much variability as possible is crucial for training a generalizable model.

Another limitation of the proposed pipeline for training a CNN in digital pathology is its reliance on the panCK tumor marker for pre-training. PanCK is an epithelium marker and is effective for tumors such as lung and breast tumors; however, for some epithelial tumors, such as pancreatic ductal adenocarcinoma (PDAC), it is not practical since panCK has similar staining in normal pancreas and PDAC. Moreover, panCK cannot be used for non-epithelial tumors. Additionally, for training a segmentation model for other tissue types (non-tumor tissues), there might not be a suitable IF or IHC marker to generate annotations for pre-training. In such situations, using a foundation model that is trained on a large dataset of pathology images might be more effective.

In this study, we included different scanners, protocols, and centers in H&E data acquisition. However, panCK imaging was done at one center by a team specialized in IF imaging with years of experience. Generating IF data at multiple labs with different protocols would be desirable. However, staining and scanning IF images at a lab with less experience or access to state-of-the-art resources might impact image quality and subsequently model pre-training. An alternative is IHC imaging, which is more stable, requires fewer resources, and is widely available, but it has its own challenges (e.g., separating tumor from non-tumor tissue).

We procured high-quality commercial samples, from tissue collection to tissue fixation and preparation. However, clinical samples (particularly from biopsies) that would be encountered when using the model for inference might not have the same quality, and the model performance might be poorer. Including low-quality tissue in test and training datasets is necessary to properly characterize such situations. Similarly, for pathologist annotations, the pathologist's experience is a major contributor to annotation quality and final model performance. These

factors were not investigated in the current study. Using data from publicly available datasets such as The Cancer Genome Atlas (TCGA) might help in assessing the impact of some of these factors on model generalizability.

Most foundation models (including the ones used here) are trained for classification tasks. However, when adapted for a segmentation task, as observed in current study, they do not provide high enough accuracy. This could be due to the fact that a large number of parameters need to be trained in the process, or the model architecture is changed. Training end-to-end foundation models for segmentation tasks has the potential to address this issue and provide improved performance that might be comparable to task-specific pre-training like the panCK pre-training used in the current study.

Another potential future direction is to use the model that was trained on NSCLC and transfer or fine-tune it on other epithelial tumors such as breast cancer, and assess model performance and determine how much extra training data would be required for such a transfer from one tumor type to another.

In conclusion a framework was proposed to determine the minimum required pathologist annotations for training a CNN for NSCLC tumor segmentation on H&E WSI. Practical study design consideration to capture NSCLC variation in H&E and challenges with each modality were also examined. The algorithm was trained using a combination of panCK and pathologist annotations, resulting in robust performance across variations in H&E images from different centers, scanners, and protocols. To account for variations in color space, color augmentation during training was found to be sufficient, and no additional improvement was achieved by using test-time color normalization. In terms of staining and imaging, using H&E and panCK on the same section yielded a slight improvement in model accuracy compared to utilizing two serial sections for H&E and panCK imaging. The use of pre-trained weights from foundation model compared to panCK pre-training and the impact of different backbone architectures in model performance were studied showing the superior performance of panCK-based pre-training. The CNN model was pre-trained on a large dataset of panCK-based tumor annotations, which helped reduce the requirement for a large dataset of pathologist annotations. Notably, even when using only 30% of the pathologist annotations, minimal reduction in model performance was observed. These findings highlight the effectiveness of the developed CNN algorithm for NSCLC tumor segmentation on H&E WSI.

## Data availability

The datasets procured and used in this study are private and not to be shared by the authors for privacy reasons. The model and code will be shared by authors upon request.

Received: 17 February 2024; Accepted: 2 August 2024
Published online: 16 September 2024

## References

1. Suster, D. I. & Mino-Kenudson, M. Molecular pathology of primary non-small cell lung cancer. *Arch. Med. Res.* **51**(8), 784–798. https://doi.org/10.1016/j.arcmed.2020.08.004 (2020).
2. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**(3), 209–249. https://doi.org/10.3322/caac.21660 (2021).
3. Brendel, M. *et al.* Weakly-supervised tumor purity prediction from frozen H&E stained slides. *eBioMedicine* **80**, 104067 (2022).
4. Bremnes, R. M. *et al.* The role of tumor stroma in cancer progression and prognosis: Emphasis on carcinoma-associated fibroblasts and non-small cell lung cancer. *J. Thorac. Oncol.* **6**(1), 209–217. https://doi.org/10.1097/JTO.0b013e3181f8a1bd (2011).
5. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z (2018).
6. Travis, W. D. *et al.* The 2015 world health organization classification of lung tumors: Impact of genetic, clinical and radiologic advances since the 2004 classification. *J. Thorac. Oncol.* **10**(9), 1243–1260. https://doi.org/10.1097/JTO.0000000000000630 (2015).
7. Nicholson, A. G. *et al.* The 2021 WHO classification of lung tumors: Impact of advances since 2015. *J. Thorac. Oncol.* **17**(3), 362–387. https://doi.org/10.1016/j.jtho.2021.11.003 (2022).
8. Khodabakhshi, Z. *et al.* Non-small cell lung carcinoma histopathological subtype phenotyping using high-dimensional multinomial multiclass CT radiomics signature. *Comput. Biol. Med.* **136**, 104752. https://doi.org/10.1016/j.compbiomed.2021.104752 (2021).
9. Kerr, K. M. & Laing, G. M. Adenocarcinoma, lung. In *Pulmonary pathology* (eds Cagle, P. T. & Kerr, K. M.) 13–38 (Springer, 2018). https://doi.org/10.1007/978-3-319-69263-0_4336.
10. Suarez, E. & Knollmann-Ritschel, B. E. C. Educational case: Squamous cell carcinoma of the lung. *Acad. Pathol.* **4**, 1–4. https://doi.org/10.1177/2374289517705950 (2017).
11. Wang, X. *et al.* Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci. Rep.* **7**(1), 13543. https://doi.org/10.1038/s41598-017-13773-7 (2017).
12. Esposito, V. *et al.* Analysis of cell cycle regulator proteins in non-small cell lung cancer. *J. Clin. Pathol.* **57**(1), 58–63. https://doi.org/10.1136/jcp.57.1.58 (2004).
13. Yatabe, Y. *et al.* Best practices recommendations for diagnostic immunohistochemistry in lung cancer. *J. Thorac. Oncol.* **14**(3), 377–407. https://doi.org/10.1016/j.jtho.2018.12.005 (2019).
14. Tokunaga, H., Teramoto, Y., Yoshizawa, A. & Bise, R. *Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology*.
15. Rączkowski, Ł *et al.* Deep learning-based tumor microenvironment segmentation is predictive of tumor mutations and patient survival in non-small-cell lung cancer. *BMC Cancer* **22**(1), 1001. https://doi.org/10.1186/s12885-022-10081-w (2022).
16. Wei, J. W. *et al.* Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci. Rep.* **9**(1), 1–8. https://doi.org/10.1038/s41598-019-40041-7 (2019).
17. Li, Z. *et al.* Deep learning methods for lung cancer segmentation in whole-slide histopathology images - the ACDC@LungHP challenge 2019. *IEEE J. Biomed. Health Inform.* **25**(2), 429–440. https://doi.org/10.1109/JBHI.2020.3039741 (2021).
18. Han, C., Pan, X. & Yan, L. *et al.* WSSS4LUAD: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma (2022).
19. Arlova, A. *et al.* Artificial intelligence-based tumor segmentation in mouse models of lung adenocarcinoma. *J. Pathol. Inform.* **13**, 100007. https://doi.org/10.1016/j.jpi.2022.100007 (2022).
20. Yang, H. *et al.* Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: A retrospective study. *BMC Med.* **19**(1), 80. https://doi.org/10.1186/s12916-021-01953-2 (2021).

21. Marini, N. *et al.* Data-driven color augmentation for H&E stained images in computational pathology. *J. Pathol. Inform.* **14**, 100183. https://doi.org/10.1016/j.jpi.2022.100183 (2023).
22. Clarke, E. L. & Treanor, D. Colour in digital pathology: A review. *Histopathology* **70**(2), 153–163. https://doi.org/10.1111/his.13079 (2017).
23. Ren, J., Hacihaliloglu, I., Singer, E. A., Foran, D. J. & Qi, X. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Medical image computing and computer assisted intervention* (eds Frangi, A. F. *et al.*) 201–209 (Springer, 2018). https://doi.org/10.1007/978-3-030-00934-2_23.
24. Riasatian, A. *et al.* Fine-tuning and training of DenseNet for histopathology image representation using TCGA diagnostic slides. *Med. Image Anal.* **70**, 102032. https://doi.org/10.1016/j.media.2021.102032 (2021).
25. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559. https://doi.org/10.1016/j.media.2022.102559 (2022).
26. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Med.* **30**(3), 863–874. https://doi.org/10.1038/s41591-024-02856-4 (2024).
27. 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).
28. Zhang, K., Guo, Y., Wang, X., Yuan, J. & Ding, Q. Multiple feature reweight DenseNet for image classification. *IEEE Access* **7**, 9872–9880. https://doi.org/10.1109/ACCESS.2018.2890127 (2019).
29. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th international conference on machine learning*. Vol. 97 (eds Chaudhuri, K., Salakhutdinov, R.). Proceedings of Machine Learning Research. PMLR, 6105–6114 (2019).
30. Tang, Y., Yang, D. & Li, W. *et al.* Self-supervised pre-training of Swin transformers for 3D medical image analysis. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 20698–20708. https://doi.org/10.1109/CVPR52688.2022.02007 (2022).
31. Weis, C. A., Weihrauch, K. R., Kriegsmann, K. & Kriegsmann, M. Unsupervised segmentation in NSCLC: How to map the output of unsupervised segmentation to meaningful histological labels by linear combination?. *Appl. Sci.* **12**(8), 3718. https://doi.org/10.3390/app12083718 (2022).
32. Kapil, A. *et al.* Domain adaptation-based deep learning for automated tumor cell (TC) scoring and survival analysis on PD-L1 stained tissue images. *IEEE Trans. Med. Imaging* **40**(9), 2513–2523. https://doi.org/10.1109/TMI.2021.3081396 (2021).
33. Brieu, N. *et al.* Automated tumour budding quantification by machine learning augments TNM staging in muscle-invasive bladder cancer prognosis. *Sci. Rep.* **9**(1), 5174. https://doi.org/10.1038/s41598-019-41595-2 (2019).
34. Bulten, W. *et al.* Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Rep.* **9**(1), 864. https://doi.org/10.1038/s41598-018-37257-4 (2019).
35. Ammeling, J. *et al.* Automated mitotic index calculation via deep learning and immunohistochemistry. In *Bildverarbeitung für die medizin 2024* (eds Maier, A. *et al.*) 123–128 (Springer, 2024).
36. Faryna, K., van der Laak, J. & Litjens, G. Tailoring automated data augmentation to H&E-stained histopathology. *Proc. Mach. Learn. Res.* **143**, 168–178 (2021).
37. Li, K. *et al.* Weakly supervised histopathology image segmentation with self-attention. *Med. Image Anal.* **86**, 102791. https://doi.org/10.1016/j.media.2023.102791 (2023).
38. Kather, J. N. *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* **16**(1), 1–22. https://doi.org/10.1371/journal.pmed.1002730 (2019).
39. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Med. Image Anal.* **67**, 101813. https://doi.org/10.1016/j.media.2020.101813 (2021).
40. Marlin, M. C. *et al.* A novel process for H&E, immunofluorescence, and imaging mass cytometry on a single slide with a concise analytics pipeline. *Cytom. Part A* **103**(12), 1010–1018. https://doi.org/10.1002/cyto.a.24789 (2023).
41. Aggarwal, A. *et al.* Intrahepatic quantification of HBV antigens in chronic hepatitis B reveals heterogeneity and treatment-mediated reductions in HBV core-positive cells. *JHEP Rep.* **5**(4), 100664. https://doi.org/10.1016/j.jhepr.2022.100664 (2023).
42. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**(1), 16878. https://doi.org/10.1038/s41598-017-17204-5 (2017).
43. Oktay, O., Schlemper, J., Folgoc, L. L. & Le, M. *et al.* Attention U-Net: Learning where to look for the pancreas. In *Medical imaging with deep learning* (2018).
44. Schlemper, J. *et al.* Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **53**, 197–207. https://doi.org/10.1016/j.media.2019.01.012 (2019).
45. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation BT - medical image computing and computer-assisted intervention–MICCAI 2015. In *Medical image computing and computer-assisted intervention* (eds Navab, N. *et al.*) 234–241 (Springer, 2015).
46. Sikaroudi, M., Hosseini, M., Gonzalez, R., Rahnamayan, S. & Tizhoosh, H. R. Generalization of vision pre-trained models for histopathology. *Sci. Rep.* **13**(1), 6065. https://doi.org/10.1038/s41598-023-33348-z (2023).
47. Zhou, Z., Lu, Q., Wang, Z. & Huang, H. Detection of micro-defects on irregular reflective surfaces based on improved faster R-CNN. *Sensors* **19**(22), 5000. https://doi.org/10.3390/s19225000 (2019).
48. Tellez, D. *et al.* Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans. Med. Imaging* **37**(9), 2126–2136. https://doi.org/10.1109/TMI.2018.2820199 (2018).
49. Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544. https://doi.org/10.1016/j.media.2019.101544 (2019).
50. Macenko, M., Niethammer, M. & Marron, J. S. *et al.* A method for normalizing histology slides for quantitative analysis. In *Proceeding of the 2009 IEEE international symposium on biomedical imaging: from nano to Macro*, *ISBI 2009*, 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250 (2009).
51. Liu, Z., Lin, Y. & Cao, Y. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In: *2021 IEEE/CVF international conference on computer vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986 (2021).
52. Cardoso, M. J., Li, W. & Brown, R. *et al.* MONAI: An open-source framework for deep learning in healthcare (2022).
53. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135 (2022).
54. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**(3), 850–862. https://doi.org/10.1038/s41591-024-02857-3 (2024).


## Acknowledgements

The authors would like to thank Dr. Metin Gurcan for reviewing the study and providing valuable input in study design and execution. We would also like to acknowledge Dr. Sangeetha Mahadevan for her help with sample procuration and providing challenging samples for testing model performance.

## Author contributions

H.M.: Conceptualization, Methodology, Software, Visualization, Data curation, Writing- Original draft preparation J.B.: Conceptualization, Writing-Reviewing and Editing, P.L., E.V., A.A.: Data curation, Writing-Reviewing and Editing, L.D.: Conceptualization, Writing-Reviewing and Editing, Supervision.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to H.M.

**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.