POLITECNICO DI TORINO
Repository ISTITUZIONALE

Shifting the Focus of Digital Pathology: The Raising Relevance of PreProcessing Phase Over Model Complexity

Original
Shifting the Focus of Digital Pathology: The Raising Relevance of PreProcessing Phase Over Model Complexity / Salvi, Massimo; Michielli, Nicola; Mogetta, Alessandro; Gambella, Alessandro; Sengur, Abdulkadir; Molinari, Filippo; Gertych, Arkadiusz. - In: IET IMAGE PROCESSING. - ISSN 1751-9659. - 20:1(2026). [10.1049/ipr2.70290]

Availability:
This version is available at: 11583/3007047 since: 2026-01-28T08:21:27Z

Publisher:
Wiley

Published
DOI:10.1049/ipr2.70290

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

24 April 2026

IET Image Processing

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Shifting the Focus of Digital Pathology: The Raising Relevance of Pre-Processing Phase Over Model Complexity

Massimo Salvi[1] | Nicola Michielli[1] | Alessandro Mogetta[1] | Alessandro Gambella[2] | Abdulkadir Sengur[3] | Filippo Molinari[1] | Arkadiusz Gertych[4,5,6]

[1]Biolab, PoliTo[BIO]Med Lab, Department of Electronics and Telecommunications, Politecnico Di Torino, Turin, Italy | [2]Pathology Unit, Department of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, Genoa, Italy | [3]Electrical-Electronics Engineering Department, Technology Faculty, Firat University, Elazig, Turkey | [4]Faculty of Biomedical Engineering, Silesian University of Technology, Gliwice, Poland | [5]Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, California, USA | [6]Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA

Correspondence: Massimo Salvi (massimo.salvi@polito.it)

Received: 16 September 2025 | **Revised:** 22 December 2025 | **Accepted:** 8 January 2026

ABSTRACT

Recent trends in computational pathology favour increasingly complex deep learning architectures, raising the question of whether such complexity is necessary for routine diagnostic tasks. This study challenges this assumption through a comprehensive analysis of the relationship between model complexity, data pre-processing, and performance across four fundamental digital pathology tasks: nuclei counting, steatosis quantification, glomeruli detection, and Ki67 proliferation index (PI) assessment. We evaluated five deep learning models of varying complexity (lightweight: MobileNetV2, U-Net, and more complex: ConvNeXt, K-Net, and Swin Transformer) combined with different image pre-processing techniques. To evaluate model performance without extensive ground truth (GT) annotations, we introduced a validation strategy utilizing the relative absolute deviation (RAD) between network predictions and correlation of performance metrics. Our findings demonstrate that pre-processing strategies, particularly stain normalization (NORM), can be more impactful than model complexity, reducing error rates by up to 50% compared to processing original (ORIG) images. With appropriate pre-processing, lightweight models achieved comparable or superior results to complex models while reducing processing times by up to 40%. Only specific tasks involving complex morphological features, such as glomeruli detection, significantly benefited from more sophisticated architectures. This study provides an evidence-based framework for selecting optimal model-pre-processing combinations in clinical settings, suggesting that investing in pre-processing pipelines rather than model complexity may be more beneficial for routine computational pathology applications.

1 | Introduction

The integration of digital technologies has dramatically transformed pathologist workflow, transitioning from traditional microscopy-based diagnostics to computational analysis of digitized glass slides [1, 2]. Furthermore, the implementation of artificial intelligence (AI), particularly deep learning models, has demonstrated significant potential in supporting procedures performed manually by pathologists that are key in diagnosis [3–5], prognostication and treatment planning [6, 7]. However, as the AI research field evolves, there is a growing tendency to develop increasingly complex deep learning models which, when reported, achieve only marginal performance improvements in basic tasks such as, for instance, nuclei segmentation or cell counting, often without considering the implications of model implementation in a clinical setting [8, 9].

The transition from microscopy-based pathology to digital pathology workflows presents significant technical challenges. While WSI systems have achieved widespread adoption in pathology

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

departments, the computational infrastructure required for advanced AI model deployment remains a significant limitation [10]. In fact, most healthcare institutions face considerable constraints in terms of GPU availability, storage capacity, and processing power, with typical pathology departments having access to standard workstations rather than high-performance computing clusters [11, 12]. These hardware limitations, combined with the need for rapid turnaround times, create practical barriers to implementing complex deep learning architectures in a clinical setting [13]. Furthermore, the cost of upgrading computational infrastructure often exceeds departmental budgets, making it crucial to develop efficient solutions that can operate within existing resources.

In recent years, we have witnessed a trend towards developing more complex deep learning models in computational pathology. These models, while achieving superior benchmark performance metrics, introduce substantial computational overhead and extend inference times [14]. This emphasis on architectural complexity has somewhat overshadowed the importance of pre-processing techniques, such as NORM and colour augmentation (CAUG), which can significantly impact model performance regardless of architectural complexity [15].

However, routine slide analysis that, for instance, involves cell counting, delineating cancer cells, or finding sclerotic glomeruli, might not require complex architectures. Our experience gained over several years in training different AI models for various segmentation tasks suggests that simpler architectures—when properly optimized and combined with appropriate pre-processing techniques—can achieve clinically acceptable accuracy [16–18]. We also noticed that a marginal difference in the intersection over union (IoU) metric, computed for the prediction masks generated by a simpler and that for a more complex model, can lead to a negligible practical benefit in WSI analysis scenarios [19].

Hence, we posit that the computational pathology workflows need a systematic evaluation of the trade-off between the model complexity and its clinical utility. Also, while maximizing the model's accuracy, its efficiency, computational complexity, and impact on the diagnostic task improvement need to be assessed as well [20].

Based on these considerations, this work aims to address the following issues in computational pathology:

- Is increased model complexity necessary for achieving clinically acceptable performance in tasks requiring WSI segmentation? This fundamental question challenges current trends in model development and has direct implications for clinical implementation. We investigated whether lightweight deep learning models could match the performance of more complex models using pipelines with image pre-processing techniques.
- What is the relative impact of data preprocessing on the accuracy of the lightweight and complex models? Understanding this relationship is crucial for developing efficient clinical solutions. To assess the impact, we examine if and how NORM and CAUG affect the performance across different models.
- What are the practical trade-offs between the model's complexity and inference time? This question directly addresses clinical workflow requirements. We analyse whether marginal improvements in the model's accuracy justify the increased computational costs by complex models.

The rest of this paper is organized as follows. Section 2 provides an exhaustive description of the proposed study, focusing on five image segmentation models of varying complexity (from lightweight MobileNetV2 to complex Swin Transformer) applied to analyse image data using three different pre-processing techniques (ORIG images, CAUG, and NORM). These models are evaluated across four distinct digital pathology tasks: nuclei counting, steatosis quantification, glomeruli detection, and Ki67 PI assessment. Key experiments include model performance evaluation on both annotated image tiles and WSIs. Experimental results are reported in Section 3, involving model testing and performance both in terms of image tiles and entire WSIs, along with comprehensive computational time analysis. Section 4 discusses the work as a whole with practical implementation recommendations for clinical settings. Finally, in Section 5, the key findings and their significance from the experimental results are briefly summarized.

2 | Materials and Methods

In this study, we trained and evaluated five image segmentation models of different complexity and used three data preprocessing techniques to assess the impact of the models on segmentation accuracy. The study design summarized in Figure 1 considers four tasks: (i) nuclei segmentation and counting in hematoxylin and eosin (H and E) stained multiorgan tissue images; (ii) steatosis quantification in H and E-stained liver tissue images; (iii) glomeruli segmentation and counting in periodic acid-Schiff (PAS) stained kidney tissue images, and (iv) Ki67 PI assessment in immunohistochemically (IHC) stained breast cancer images. Model evaluation was carried out in two phases: using image tiles with pathologist GT annotations (Figure 1a) and using WSIs to simulate analysis in a clinical setting (Figure 1b). In the first phase, stain perturbation was also applied to test image tiles to simulate staining variability in slides from multiple centres. Both analyses were quantitatively assessed with segmentation-based and task-specific performance metrics to comprehensively evaluate two key aspects: the impact of model complexity and the relative importance of data pre-processing techniques.

2.1 | Datasets

This study utilized four datasets, each corresponding to a specific segmentation task in computational pathology. Each dataset included image tiles and WSIs organized in subsets as follows:

i. ORIG training, validation and test image tiles with pathologist GT annotations;
ii. stain-perturbed test image tiles with pathologist GT annotations to simulate staining variability;
iii. WSIs without GT annotations.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

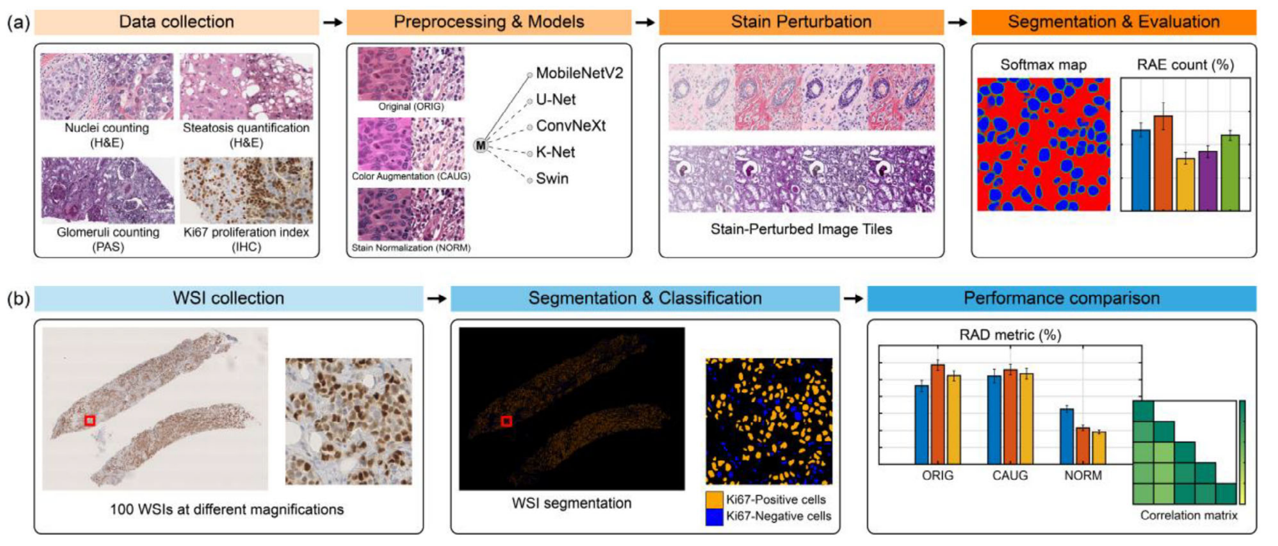

FIGURE 1 | Overview of the experimental design. (a) Five segmentation models were applied to analyse image data pre-processed using ORIG, CAUG, and NORM techniques, and evaluated across four different digital pathology tasks. Stain perturbation was applied to simulate staining variability typically encountered in multi-centre datasets. Performance assessment included both segmentation-based and task-specific metrics evaluation using the relative absolute error (RAE); and (b) WSI processing to reproduce slide analysis in a clinical setting: an example of breast tissue WSI with segmentation of Ki67-positive and Ki67-negative cell nuclei to evaluate the Ki67 PI. Correlation matrix and RAD metric were computed for performance comparison.

These datasets, which include a variety of tissue types and different staining colorations, were gathered from different repositories, and, when considered together, were used to comprehensively evaluate model performance in different quantitative tasks in pathology. Dataset characteristics are provided in Table 1.

2.1.1 | ORIG Image Tiles

For nuclei segmentation and counting, we employed the MoNuSeg dataset [21], which comprises 30 H and E-stained tiles from tissue samples of various organs (i.e., breast, liver, kidney, prostate, bladder, colon, and stomach) with overall 21,713 nuclei with GT annotations. Captured at 40x magnification, this dataset represents a variety of cell types, nuclear phenotypes, and cell confluency, enabling development of nuclear instance segmentation models that can also be used for nuclei counting.

The HEPASS dataset [22] was used for steatosis quantification. It consists of 560 H and E-stained liver tissue tiles acquired at 20x magnification, with 10,952 lipid vesicles hand-annotated by a pathologist. This dataset can be used to assess accuracy of semantic segmentation models in identifying and delineating lipid vesicles in liver specimens.

For glomeruli segmentation and counting, we used the RENTAG dataset [16], which includes 610 PAS-stained kidney tissue tiles captured at 10x magnification with 587 hand annotated glomeruli. This dataset allows for the evaluation of instance segmentation and counting of healthy and pathological glomeruli—a task frequently performed by nephropathologists.

Finally, for the PI assessment in IHC stained images, we employed a private dataset of Ki67-stained breast tissue images, consisting of 179 image tiles captured at 20x magnification, with 38,734 and 44,553 cell nuclei manually delineated and labelled as Ki67-positive or Ki67-negative, respectively.

2.1.2 | Stain-Perturbed Test Image Tiles

Coloration of tissue staining in images of glass slides can vary due to differences in staining protocols and slide scanner settings. To simulate the effect of stain colour changes and variability in image appearance of the same slide scanned using different devices, we perturbed the HSV colour channels in our test image tiles four times; that is, for each test tile, 4 new image tiles with altered staining coloration were generated [23, 24]. The stain perturbation process involved converting the ORIG RGB image into the HSV colour space, modifying the saturation and value channels while preserving the hue component, and converting the image with perturbed colour channels back to the RGB space. The saturation and value channels were modified by respectively increasing and decreasing their contrast by 50%. Figure 2 shows example images from each of the four datasets along with their corresponding stain-perturbed variants. The numbers of stain-perturbed image tiles for model testing are summarized in Table 1.

2.1.3 | WSIs

For each task we collected separate sets, 100 WSIs each (400 in total). The set for evaluating nuclei segmentation and counting consists of H and E-stained slides of breast, colon, lung, and prostate tissue, including 66 WSIs sourced from The Cancer Genome Atlas (TCGA) archive (https://portal.gdc.cancer.gov) and 34 private WSIs acquired at the Michele and Pietro Ferrero Hospital (Cuneo, Italy) and A.O.U. Città Della Salute e Della Scienza Hospital (Turin, Italy). Specifically, there were

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by Cochraneltalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Dataset characteristics for the four digital pathology tasks. For each task, the number of ORIG image tiles in training, validation and test sets is reported, along with scanning objective magnification and pixel size. The number of stain-perturbed test image tiles and WSIs used for evaluation is also indicated.

Task (stain)	Organ	Tiles and WSIs (Subset)	Quantity	Pixel size (magnification)	Datasets (total quantity)
Nuclei counting (H and E)	Multiorgan	Tiles (train)	13	0.25 µm (40x)	MoNuSeg [21] (30)
		Tiles (validation)	3		
		Tiles (test)	14		
		Tiles (perturbed test)	56		
		WSIs (test)	100	0.25 µm (40x) 0.50 µm (20x)	TCGA (66) Private (34)
Steatosis quantification (H and E)	Liver	Tiles (train)	454	0.47 µm (20x)	HEPASS [22] (560)
		Tiles (validation)	56		
		Tiles (test)	50		
		Tiles (perturbed test)	200		
		WSIs (test)	100	0.25 µm (40x) 0.50 µm (20x)	TCGA (50) Private (50)
Glomeruli counting (PAS)	Kidney	Tiles (train)	430	0.93 µm (10x)	RENTAG [16] (610)
		Tiles (validation)	70		
		Tiles (test)	110		
		Tiles (perturbed test)	440		
		WSIs (test)	100	0.50 µm (20x)	Private (100)
Ki67 proliferation index assessment (IHC)	Breast	Tiles (train)	126	0.44 µm (20x)	Private (179)
		Tiles (validation)	25		
		Tiles (test)	28		
		Tiles (perturbed test)	112		
		WSIs (test)	100	0.22 µm (40x) 0.44 µm (20x)	Private (100)

17 (private) and 8 (TCGA) WSIs with breast specimens, 25 (TCGA) WSIs with colon specimens, 25 (TCGA) WSIs with lung specimens, and 17 (private) and 8 (TCGA) slides with prostate specimens. For steatosis quantification, 50 H and E-stained liver WSIs were sourced from TCGA and 50 private WSIs from A.O.U. Città Della Salute e Della Scienza Hospital (Turin, Italy). Both, the 100 PAS-stained kidney WSIs used for glomeruli counting and the 100 Ki67-stained breast tissue WSIs used for Ki67 PI assessment were sourced privately from the Michele and Pietro Ferrero Hospital (Cuneo, Italy) and A.O.U. Città Della Salute e Della Scienza Hospital (Turin, Italy). All WSIs were digitized with 20x or 40x scanning object magnification (Table 1).

2.2 | Model Architecture and Training

In this study, we compared five distinct deep learning model architectures, both lightweight and complex. These models were selected based on prior research to evaluate the relationship between model complexity, performance, and computational efficiency. Table 2 summarizes the key characteristics of each model, ordered based on model complexity.

The MobileNetV2 [25] represents a lightweight convolutional neural network (CNN) with depth-wise separable convolutions designed for computational efficiency. The U-Net model [26] comprises a standard encoder-decoder architecture with skip connections. The ConvNeXt model [27] introduces an approach that combines CNN with vision transformer (ViT) characteristics. The K-Net model [28] further increases complexity by implementing kernel updates and interaction mechanisms. The Swin Transformer [29], which is the most complex of the five models, represents a ViT architecture that employs a hierarchical structure with shifted windows to efficiently perform local self-attention. More details about the model architectures can be found in Supplementary Material (Table S1).

For model training, the training and validation image tiles (Table 1) were split into smaller patches with the patch size selected for each task as follows:

- Nuclei segmentation and counting (H and E): 320 × 320 pixels, 2-class task (nuclei instances vs background);
- Steatosis quantification (H and E): 416 × 416 pixels, 2-class task (steatosis vs background);

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

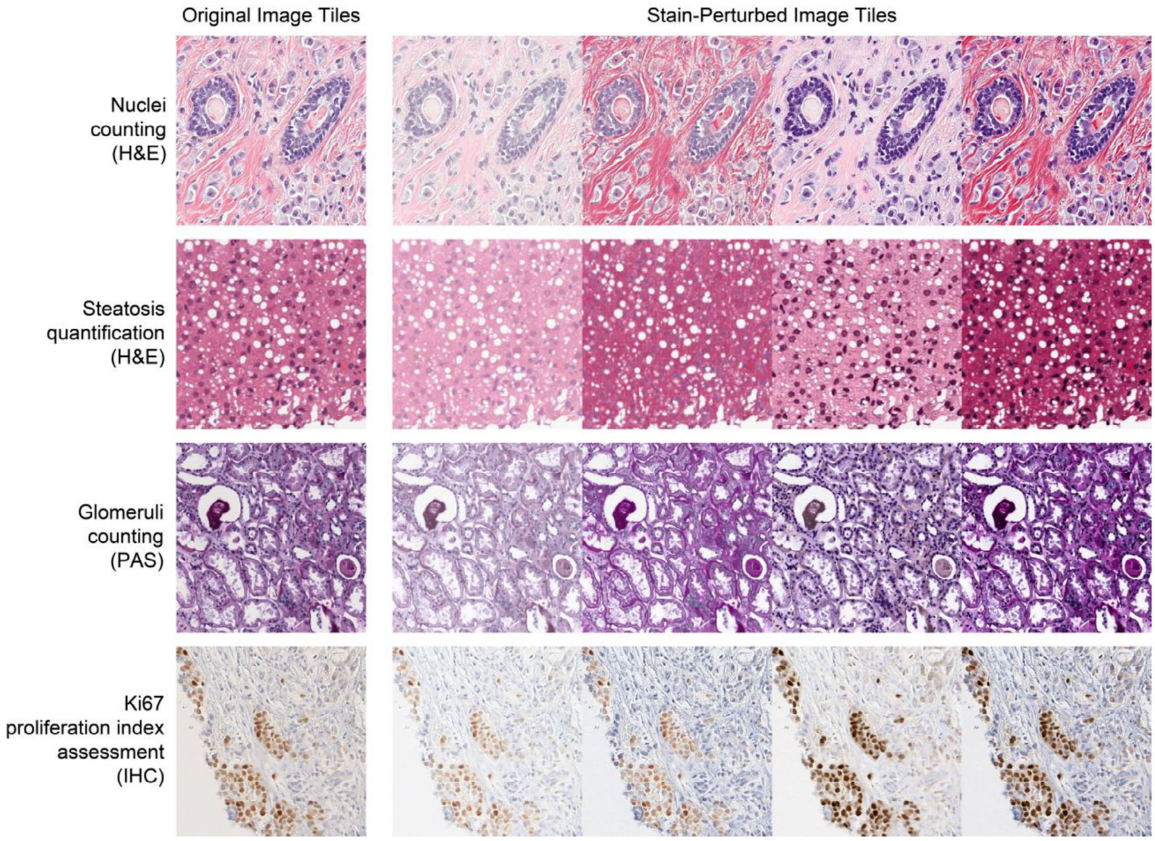

FIGURE 2 | Examples of ORIG image tiles (first column) and stain-perturbed image tiles (from second to fifth column).

TABLE 2 | Main characteristics of the deep learning models used in the study.

Model	Type	Parameters (millions)	Complexity (GFLOPs)	Total size (MB)
MobileNetV2 [25]	Lightweight CNN	9.7	2.0	37.1
U-Net [26]	Simple CNN	28.9	8.6	110
ConvNeXt [27]	Advanced CNN	60.13	25.1	229
K-Net [28]	Deep CNN	79.6	35.2	303.9
Swin [29]	ViT	121.2	19.9	462

- Glomeruli segmentation and counting (PAS): 512 × 512 pixels, 3-class task (healthy glomeruli vs pathological glomeruli vs background);
- Ki67 PI assessment (IHC): 320 × 320 pixels, 3-class task (Ki67-negative cells vs Ki67-positive cells vs background).

Model training was performed using datasets described in Table 1. For each task, all five models (MobileNetV2, U-Net, ConvNeXt, K-Net, and Swin) were trained using identical training sets of image tiles to ensure a fair comparison of model performances. For steatosis quantification (H and E) and glomeruli segmentation and counting (PAS), the training and validation image tiles were used, since the receptive field of each model matched the image tile size. For nuclei segmentation and counting (H and E), we prepared 468 patches for training and 108 patches for validation from the image tiles. For Ki67 PI assessment (IHC), 1434 patches were used for training and 270 for validation. During training, the on-the-fly data augmentation was applied, including rotations and flips. All models were trained using Adam optimizer with Dice score loss, which is particularly suited for medical image segmentation tasks as it directly optimizes spatial overlap between predicted and GT segmentations. The Dice loss function is defined as:

$$Dice_{loss} = 1 - \frac{2 \cdot |P \cap G| + \varepsilon}{|P| + |G| + \varepsilon} \tag{1}$$

where P represents the predicted segmentation mask and G represents the GT mask. $|P \cap G|$ denotes the intersection between prediction and GT, while $|P|$ and $|G|$ represent their respective cardinalities. ε is a small constant added to avoid division by zero. This loss function ranges from 0 to 1, where 0 indicates perfect overlap between prediction and GT. The loss was monitored during 50 epochs of training with an initial learning rate of 1e-4 and stepwise learning rate decay, and a batch size of 4 tiles.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

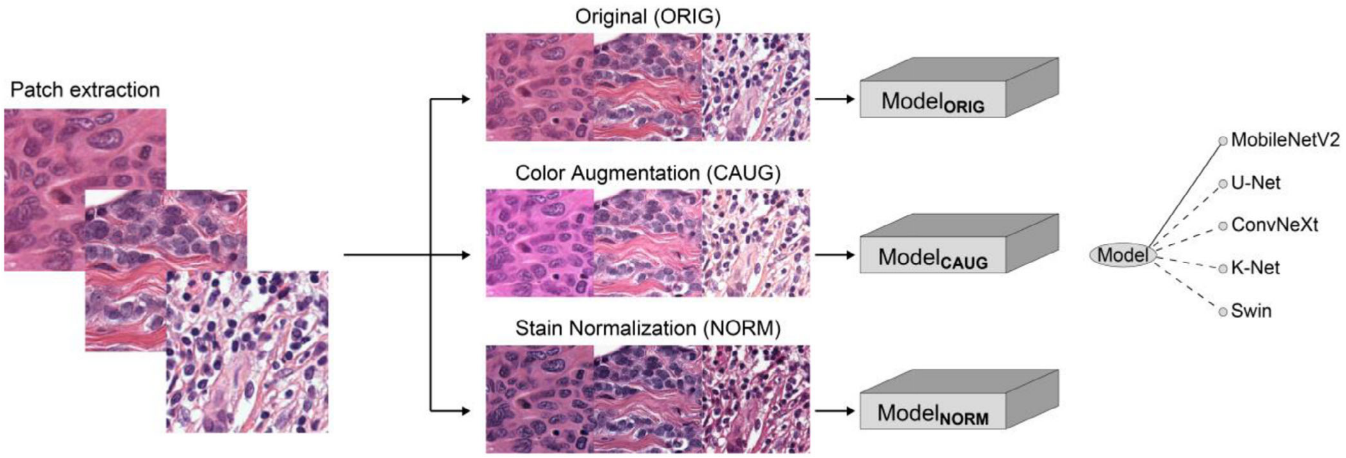

FIGURE 3 | Examples of H and E-stained patches subjected to the three data preprocessing techniques (ORIG, CAUG, NORM) used during inference by five different deep learning models (MobileNetV2, U-Net, ConvNeXt, K-Net, and Swin).

To assess the impact of data preprocessing techniques on model performance, we trained each model under three different scenarios (Figure 3). In the first scenario, only ORIG patches were used in training. In the second scenario, CAUG was applied to each patch by randomly modifying the hue, saturation, and brightness of the images. Specifically, the hue was shifted by a random value between −0.05 and 0.05, while saturation and brightness were scaled by a random factor between 0.8 and 1.2. This process allowed simulating various staining variations that are encountered in real-world data, providing the model with a more diverse and representative set of images. In the third scenario, NORM was applied using a previously developed tool [30]. In contrast to CAUG, NORM aims to standardize the color appearance of the images, thereby reducing the variability in staining patterns across different samples [31].

In total, 15 deep learning models were trained; each of the five models listed in Table 2 was trained de novo using the ORIG, CAUG, and NORM image data. This approach allowed us to evaluate the impact of deep learning model complexity and data pre-processing techniques on the segmentation accuracy and inference times.

2.3 | Task Description

We focused on four distinct tasks in digital pathology, each with its own specific requirements and clinical relevance (Figure 4):

- Nuclei segmentation and counting (H and E): performed in H and E-stained images is crucial for various pathological assessments, including nucleus-based cell profiling, tumor grading, tumor proliferation, and cellularity assessment [32, 33]. This task involves instance segmentation to identify individual nuclei and is challenging due to cell clustering, overlapping, and variabilities in nuclear shape and size.
- Steatosis quantification (H and E) in H and E-stained liver tissue is essential for assessing the severity of fatty liver disease [17]. It requires semantic segmentation to identify and measure the area of lipid vesicles which can vary in size and confluency. According to Liquori et al. [34], measuring the vesicle area is crucial in the assessment of steatosis, which, based on the vesicle area, is classified as macrovesicular (MACRO) steatosis (vesicle area > 175 μm^2) or microvesicular (MICRO) steatosis (vesicle area ≤ 175 μm^2).
- Glomeruli segmentation and counting (PAS) is vital for assessing renal function and diagnosing kidney diseases [35]. This task involves instance segmentation, identifying and delineating healthy and pathological glomeruli, which can be challenging due to their complex structure and pathological changes that can affect glomerular morphology.
- Ki67 PI assessment (IHC): determining this index, defined as the ratio of Ki67-positive tumour cell nuclei to the total tumor cell nuclei in IHC stained specimen tissue is crucial for assessing tumour aggressiveness, outcome prediction and treatment planning in breast cancer [36]. This task requires nuclei detection, segmentation, and quantification of marker intensity inside the detected nuclei to determine tumor cells that are Ki67-positive or Ki67-negative.

2.4 | Inference and Evaluation Metrics

During network inference, we applied classical post processing techniques, including thresholding on the softmax output to obtain a prediction mask, followed by mask refinement with morphological operators. All task-specific post processing steps are provided in the Supplementary Material (Figure S1). For the WSI analysis, we first applied a tissue detection algorithm [17], both for H and E, PAS, and IHC stained slides, to identify tissue area for analysis and exclude background, and subsequently applied the models only to the tissue area. An example WSI segmentation is provided in the Supplementary Material (Figure S2).

To comprehensively evaluate model performance, we employed both pixel-based metrics and task-specific segmentation accuracy metrics (Table 3). For pixel-based evaluations, the precision, recall, and IoU were used. These metrics were computed for all detected instances disregarding the class type, that is, pixels

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

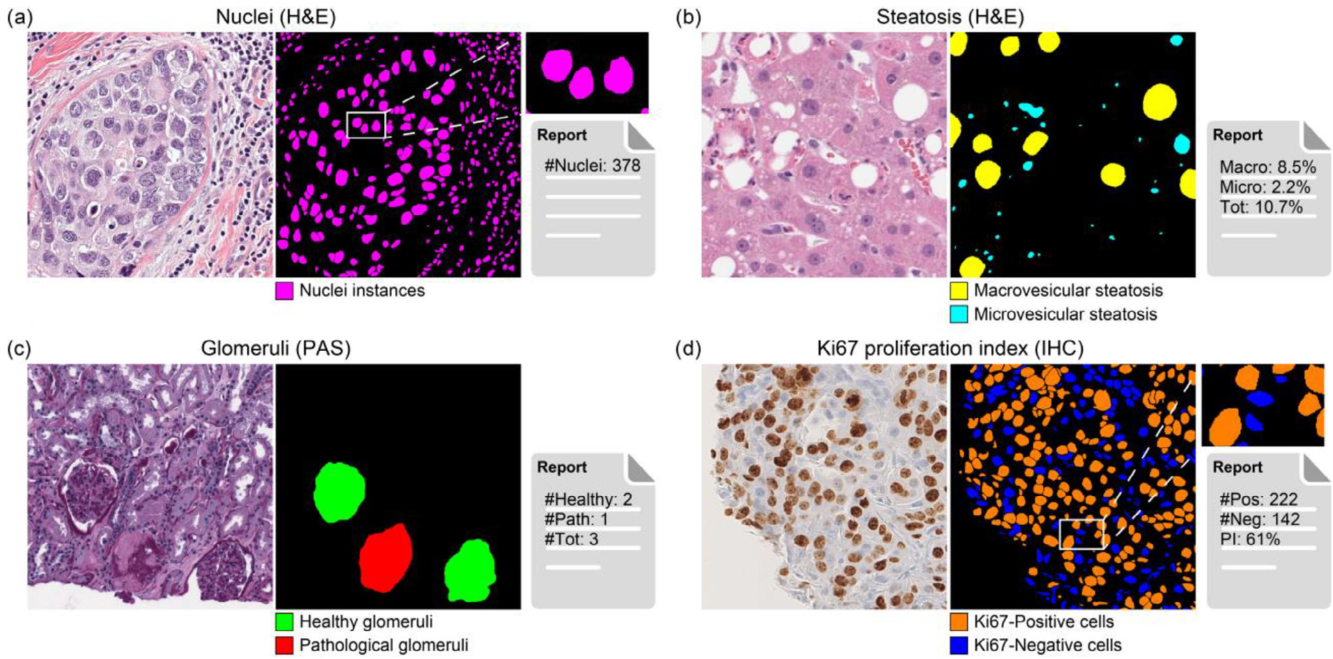

FIGURE 4 | Examples for segmentation tasks evaluated in the study. (a) nuclei counting in H and E-stained tissue; (b) steatosis quantification in H and E-stained liver tissue; (c) glomeruli counting in PAS-stained kidney tissue; and (d) Ki67 PI assessment in IHC stained breast cancer tissue. Each panel shows the ORIG image, prediction mask with segmented and classified cellular structures, and an example quantification result.

TABLE 3 | Overview of segmentation tasks and evaluation metrics. IoU: intersection over union; $Dice_{OBJ}$: Object-based Dice index; AJI: aggregated Jaccard Index; RAE_{COUNT}: relative absolute error for nuclei count; AE_{MACRO}: absolute error for macro vesicular steatosis; AE_{MICRO}: absolute error for micro vesicular steatosis; AE_{TOT}: absolute error for total steatosis; $AE_{HEALTHY}$: absolute error for healthy glomeruli count; AE_{PATH}: absolute error for pathological glomeruli count; $\mathrm{AE_{TOT}}$: absolute error for total glomeruli count; RAE_{POS}: relative absolute error for Ki67-positive cells count; RAE_{NEG}: relative absolute error for Ki67-negative cells count; and AE_{PI}: absolute error for Ki67 PI.

Task (Stain)	Segmentation type	Pixel-based metrics	Task-specific metrics
Nuclei (H and E)	Instance segmentation	Precision, Recall, IoU	$\mathrm{Dice_{OBJ}}$, AJI, $\mathrm{RAE_{COUNT}}$
Steatosis (H and E)	Semantic segmentation		$\mathrm{AE_{MACRO}}$, $\mathrm{AE_{MICRO}}$, $\mathrm{AE_{TOT}}$
Glomeruli (PAS)	Instance segmentation		$\mathrm{AE_{HEALTHY}}$, $\mathrm{AE_{PATH}}$, $\mathrm{AE_{TOT}}$
Ki67 proliferation index (IHC)	Instance segmentation		$\mathrm{RAE_{POS}}$, $\mathrm{RAE_{NEG}}$, $\mathrm{AE_{PI}}$

with predicted classes other than background were merged into one. In addition to pixel-based metrics, we employed task-specific metrics including the absolute error (AE) defined as the absolute difference between GT and predicted values, and the RAE defined as the ratio of the AE and the GT value. For each task, the task-specific metrics are detailed as follows:

- Nuclei segmentation and counting: we used the object-based Dice index ($Dice_{OBJ}$), which evaluates the object detection quality using the harmonic mean of the object-based precision and recall metrics. We also employed the aggregated Jaccard index (AJI) [37] to measure the quality of instance segmentation by overlaying the predicted and GT object instances. Additionally, we calculated the RAE_{COUNT} to compare GT and automatic nuclei counts.
- Steatosis quantification: we calculated the AE in terms of relative area for MACRO, MICRO, and total steatosis. The AE quantifies the AE between the percentage area occupied by steatosis in the predicted and GT segmentation masks for each vesicle category.
- Glomeruli segmentation and counting: for this task, we computed the AE separately for healthy, pathological, and total glomeruli counts to quantify the difference between the predicted and GT counts of each glomerular type.
- Ki67 pi assessment: we used the RAE for Ki67-positive and Ki67-negative nuclei count to assess the difference between the predicted and GT counts. Additionally, we calculated the AE_{PI} for the Ki67 PI, defined as the ratio of the Ki67-positive nuclei and the total number of tumour nuclei.

Taken together, the pixel-based and task-specific metrics allowed for evaluating the performance of the five models and assessing their usefulness in digital pathology. Paired-sample t-tests were also carried out to identify differences and potential improvements in the evaluation metrics yielded by the models trained using patches processed by different preprocessing techniques.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by Cochraneltalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To evaluate model performance at the WSI level, due to the lack of WSI-level GT, our primary goal was to quantitatively assess the consistency between models, rather than the accuracy evaluated using GT. For each task, we used the Pearson correlation coefficient (PCC) with paired-sample t-tests to evaluate the relationship between the task-specific outputs of all possible pairs of deep learning models used, and the proposed RAD metric to measure the deviation in the task-specific outputs of all deep learning networks with respect to a central average value (used in the absence of GT). Mathematically, the RAD metric for each WSI pre-processed using the ORIG, CAUG, or NORM technique is defined as:

$$RAD_p\,(\%) = \frac{1}{5}\sum_{i=1}^{5} 100 \cdot \frac{|y_i - \bar{y}|}{\bar{y}};\ p = \begin{cases} 1: ORIG \\ 2: CAUG \\ 3: NORM \end{cases} \quad (2)$$

where y_i denotes the output of the i-th deep learning model preprocessed using the p-th technique and $\bar{y}$ is the average value of the outputs of the five models preprocessed using the same p-th technique. More specifically:

- For nuclei counting, y_i and $\bar{y}$ refer to the total number of nuclei.
- For steatosis quantification, three types of RAD can be computed, hence y_i and $\bar{y}$ refer to the relative area occupied by MACRO, MICRO, and total steatosis, respectively.
- For glomeruli counting, three types of RAD can be computed, hence y_i and $\bar{y}$ refer to the number of healthy, pathological and total glomeruli, respectively.
- For Ki67 PI assessment, three types of RAD can be computed, hence y_i and $\bar{y}$ refer to the Ki67-positive and Ki67-negative nuclei count, and the Ki67 PI, respectively.

Additionally, we measured WSI processing times by each deep learning model.

2.5 | Experimental Setup

Our experimental setup was designed to reflect both state-of-the-art research capabilities and practical clinical scenarios. We trained and tested all models in Python (v. 3.8.19) using PyTorch (v2.1.0) with the MM segmentation library (v1.2.2) [38] from the OpenMMLab framework. For model training, we employed a high-performance workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM), an Intel Core i9-14900K CPU and 256 GB of RAM. To better simulate the computational capability of hardware typically available in a clinical setting, we conducted model testing and evaluation using a lower performance system. This setup comprised an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM), an Intel Core i5-13600K CPU and 64 GB of RAM.

3 | Results

This section presents accuracy metrics gathered from numerical experiments involving model testing using the original test image tiles, stain-perturbed test image tiles, and WSIs. Since the test image tiles (both the ORIG and stain-perturbed ones) had GT annotations available, the pixel-based and task-specific metrics were computed. For experiments with the WSIs, GT annotations were not available, hence we used the PCC and paired-sample t-tests, and the RAD metric (Equation 1) to compare quantitative redouts derived from each model output for each of the four segmentation tasks. The WSI-based analysis also includes comparisons of WSI processing times. As outlined in Section 2 (Table 2), five deep learning models of various complexity, i.e., MobileNetV2 (lightweight CNN), U-Net (simple CNN), ConvNeXt (advanced CNN), K-Net (deep CNN), and Swin (ViT) were tested.

3.1 | Performance Evaluation Using ORIG Test Image Tiles

Table 4 presents the pixel-based metric (IoU) and the task-specific metric (AE or RAE) across all tasks. For nuclei counting, we report RAE_{COUNT} between the manual and automatic nuclear counts. For steatosis quantification, $\mathrm{AE}_{\mathrm{TOT}}$ represents the difference between the predicted and manual percentage areas of total steatosis. Glomeruli counting uses AE_{TOT} between manual and automatic counts, while Ki67 assessment employs AE_{PI} between manual and automatic proliferation indices. Complete metrics are available in the supplementary material (Tables S2–S5).

In nuclei segmentation and counting, IoU values remained stable around 0.674, while RAE_{COUNT} showed a higher variation from 7.62% to 15.4%. ConvNeXt demonstrated consistent performance, achieving the lowest RAE_{COUNT} (7.68% with CAUG and 7.62% with NORM).

Steatosis quantification presented the lowest IoUs (average 0.443) among all tasks, indicating challenging lipid vesicle segmentation regardless of the model or the pre-processing method applied. However, the average AE_{TOT} of 0.648% suggests excellent vesicle detection capability across all models. ConvNeXt with CAUG achieved the highest segmentation accuracy (IoU = 0.479), followed closely by U-Net with NORM (IoU = 0.475). Notably, U-Net showed the lowest AE_{TOT} values when trained on the ORIG (0.338%) and stain normalized images (0.452%).

All models demonstrated robust glomerular segmentation and counting performance, with Swin achieving the highest average accuracies (IoU = 0.892, AE_{TOT} = 0.064). U-Net performed slightly lower but maintained consistent results (IoU = 0.852, AE_{TOT} = 0.158) regardless of the preprocessing method used.

For Ki67 proliferation assessment, MobileNetV2 showed lower average performance (IoU = 0.611, AE_{PI} = 5.08%) compared to other models. U-Net, ConvNeXt, K-Net, and Swin achieved higher average IoUs (0.646, 0.658, 0.644, and 0.645, respectively) and lower average AE_{PI} values (3.90%, 3.77%, 3.79%, and 3.61%, respectively). Swin with CAUG pre-processing, achieved the lowest AE_{PI} at 3.30%.

3.2 | Performance Evaluation Using Stain-Perturbed Test Image Tiles

To simulate stain variability in histological images, we applied the stain perturbation procedure to the ORIG test image tiles

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Pixel-based and task-specific metrics computed on ORIG test image tiles (average value) for all segmentation tasks in digital pathology. Arrows denote direction of metric improvement and the best values for each metric are highlighted in bold. IoU: intersection over union; RAE_{COUNT}: relative absolute error for nuclei count; AE_{TOT}: absolute error for total steatosis; AE_{TOT}: absolute error for total glomeruli count; and AE_{PI}: absolute error for PI.

		Nuclei (H and E)		Steatosis (H and E)		Glomeruli (PAS)		Ki67 proliferation index (IHC)	
Pre-processing	**Models**	**IoU ↑**	**RAE_{COUNT} (%) ↓**	**IoU ↑**	**AE_{TOT} (%) ↓**	**IoU ↑**	**AE_{TOT} ↓**	**IoU ↑**	**AE_{PI} (%) ↓**
Original images	MobileNetV2	0.659	15.4	0.415	0.866	0.872	0.091	0.608	5.02
	U-Net	0.690	8.62	0.447	**0.338**	0.847	0.100	0.647	3.77
	ConvNeXt	0.687	8.63	0.461	0.475	0.886	0.073	**0.661**	3.50
	K-Net	0.675	8.63	0.455	0.623	0.886	0.064	0.642	3.76
	Swin	0.685	8.82	0.430	0.802	**0.902**	**0.036**	0.650	3.38
Color augmented images	MobileNetV2	0.651	13.3	0.367	1.170	0.892	0.055	0.619	4.75
	U-Net	0.672	11.5	0.471	0.514	0.841	0.218	0.656	4.04
	ConvNeXt	0.676	7.68	**0.479**	0.726	0.876	0.091	0.656	3.89
	K-Net	0.678	9.29	0.441	0.619	0.896	0.073	0.651	3.44
	Swin	0.675	8.75	0.459	0.675	0.884	0.091	0.646	**3.30**
Stain normalized images	MobileNetV2	0.646	12.2	0.400	0.653	0.889	0.082	0.607	5.48
	U-Net	0.671	11.8	0.475	0.452	0.867	0.155	0.635	3.88
	ConvNeXt	0.675	**7.62**	0.435	0.622	0.872	0.064	0.656	3.93
	K-Net	0.671	9.01	0.455	0.542	0.875	0.091	0.638	4.17
	Swin	**0.692**	12.2	0.453	0.649	0.890	0.064	0.638	4.15

and reassessed performances of the models using these stain-perturbed image tiles. The goal was to assess the impact of the simulated staining variability on model's performance trained using the ORIG, CAUG and stain normalized images. IoUs, AEs, and RAEs specific for each segmentation task are reported in Table 5. Additional results from these experiments can be found in supplementary material (Tables S6–S9).

Table 5 shows that NORM improved both IoU and task-specific metrics (AEs and RAEs) across all tasks, except for glomeruli counting, where performance between the ORIG and normalized images was comparable. To prove the significance of this improvement, paired-sample t-tests were conducted to compare the best configuration model for ORIG vs. stain normalized data preprocessing. For both H and E-stained tasks, NORM led to statistically significant improvements ($p < 0.001$), reducing the RAE_{COUNT} in nuclei counting and the AE_{TOT} in steatosis quantification. In PAS-stained images, glomeruli counting showed no significant improvement with NORM, which was an expected outcome given the similar error values between preprocessing approaches. The most pronounced effect was observed in IHC stained images, where Ki67 PI assessment demonstrated a highly significant reduction in the AE_{PI} ($p < 0.001$) under NORM preprocessing.

Given these findings, we compared performance metrics of the five models trained using the stain normalized data preprocessing and tested using the stain-perturbed image tiles to assess each model's utility in counting of nuclei and quantification of vesicles in H and E, glomeruli counting in PAS, and Ki67 PI assessment in IHC stained images (Figure 5).

The results for pixel-based IoU metric showed comparable values across models, except for steatosis quantification, where the IoU fluctuated in the 0.381–0.450 range. Regarding the task-specific metrics, the results for nuclei counting showed variations between network outputs with a RAE_{COUNT} ranging from 7.93% (ConvNeXt) to 14.3% (U-Net). For total steatosis quantification, the AE_{TOT} values remained within a narrow range (0.523–0.719%), with similar consistency observed in Ki67 PI assessment (AE_{PI} range: 5.09–6.87%). In total glomeruli counting, four of the five models demonstrated comparable AE_{TOT} values (0.084–0.107), while U-Net showed a notably higher AE_{TOT} of 0.377.

3.3 | Performance Comparison on WSIs

Although tile-level analysis offers useful information, differences observed in a limited number of tiles may either disappear or become more pronounced when evaluating the entire WSIs. This motivates the importance of extending our evaluation to the WSI level, where the interplay between model architecture and pre-processing methods can be observed across larger, more representative tissue areas. However, since manual annotation of entire WSIs is hard to collect and GT data at the WSI level are often not available, we focused our analysis on measuring the correlation between results yielded by different models when combined with different pre-processing approaches.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 | Pixel-based and task-specific metrics computed on stain-perturbed test image tiles (average value) for all segmentation tasks in digital pathology. Arrows denote direction of metric improvement and the best values for each metric are highlighted in bold. IoU: intersection over union; RAE_{COUNT}: relative absolute error for nuclei count; AE_{TOT}: absolute error for total steatosis; AE_{TOT}: absolute error for total glomeruli count; and AE_{PI}: absolute error for PI.

		Nuclei (H and E)		Steatosis (H and E)		Glomeruli (PAS)		Ki67 proliferation index (IHC)	
Pre-processing	**Models**	**IoU ↑**	**RAE_{COUNT} (%) ↓**	**IoU ↑**	**AE_{TOT} (%) ↓**	**IoU ↑**	**AE_{TOT} ↓**	**IoU ↑**	**AE_{PI} (%) ↓**
Original images	MobileNetV2	0.608	17.9	0.314	1.320	0.723	0.582	0.511	19.9
	U-Net	0.614	15.2	0.391	0.671	0.633	1.210	0.561	11.6
	ConvNeXt	0.650	13.7	0.395	0.734	0.851	0.168	0.587	8.56
	K-Net	0.630	12.5	0.398	0.759	0.840	0.295	0.577	9.39
	Swin	0.660	12.0	0.364	0.926	**0.883**	0.096	0.577	8.57
Color augmented images	MobileNetV2	0.615	14.6	0.307	1.460	0.727	0.664	0.535	12.8
	U-Net	0.610	14.7	0.402	0.806	0.718	1.140	0.582	10.4
	ConvNeXt	0.651	12.2	0.426	0.813	0.849	0.193	0.587	8.34
	K-Net	0.653	11.3	0.398	0.696	0.854	0.491	0.584	7.32
	Swin	0.653	11.0	0.413	0.784	0.873	0.125	0.573	9.33
Stain normalized images	MobileNetV2	0.640	12.2	0.381	0.719	0.871	0.093	0.562	6.87
	U-Net	0.658	14.3	**0.450**	**0.523**	0.812	0.377	0.576	5.93
	ConvNeXt	0.670	**7.93**	0.419	0.668	0.867	0.091	**0.607**	5.43
	K-Net	0.667	8.94	0.446	0.595	0.860	0.107	0.591	**5.09**
	Swin	**0.688**	11.4	0.423	0.717	0.880	**0.084**	0.589	5.46

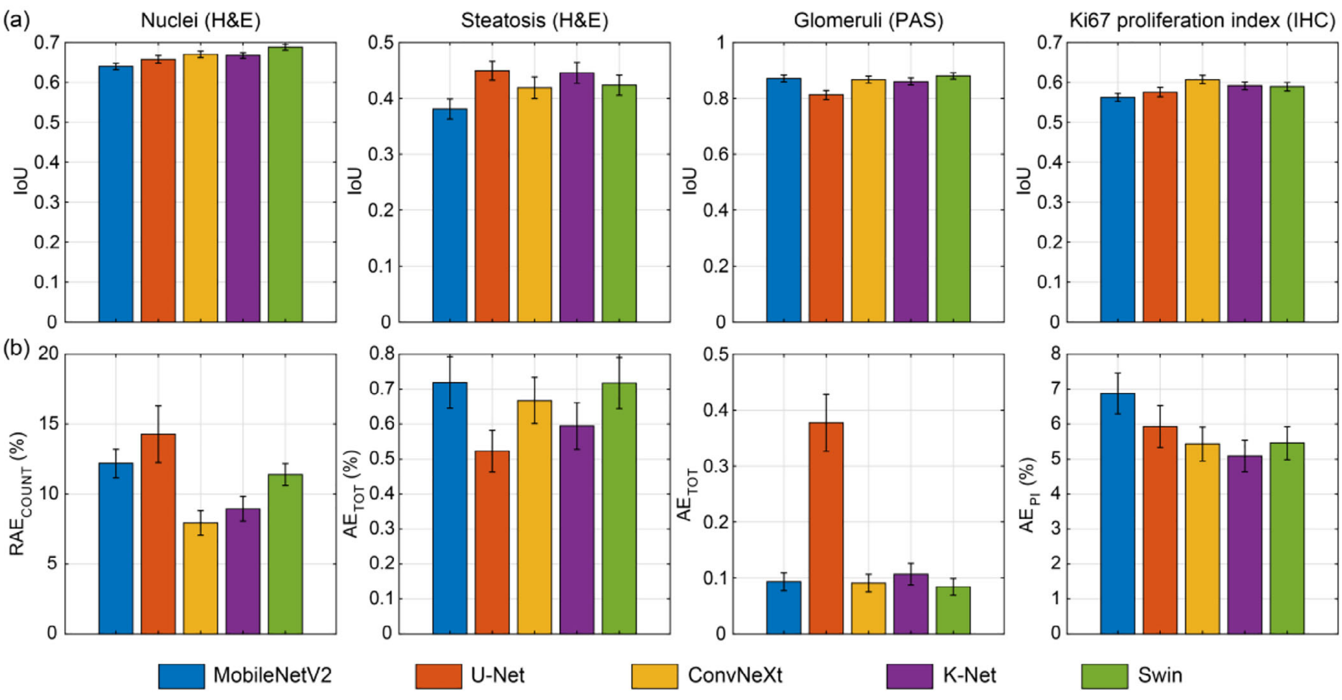

FIGURE 5 | Pixel-based IoU metric (a) and task-specific error metrics; (b) comparison of the five models trained using stain normalized data pre-processing technique and benchmarked using stain-perturbed test image tiles across four segmentation tasks. The bar height and whiskers represent the mean ± standard error. RAE_{COUNT} (first column), AE_{TOT} (second column), AE_{TOT} (third column), and AE_{PI} (fourth column), represent the nuclear count error, total steatosis percentage area error, total glomeruli count error, and Ki67 PI error, respectively.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

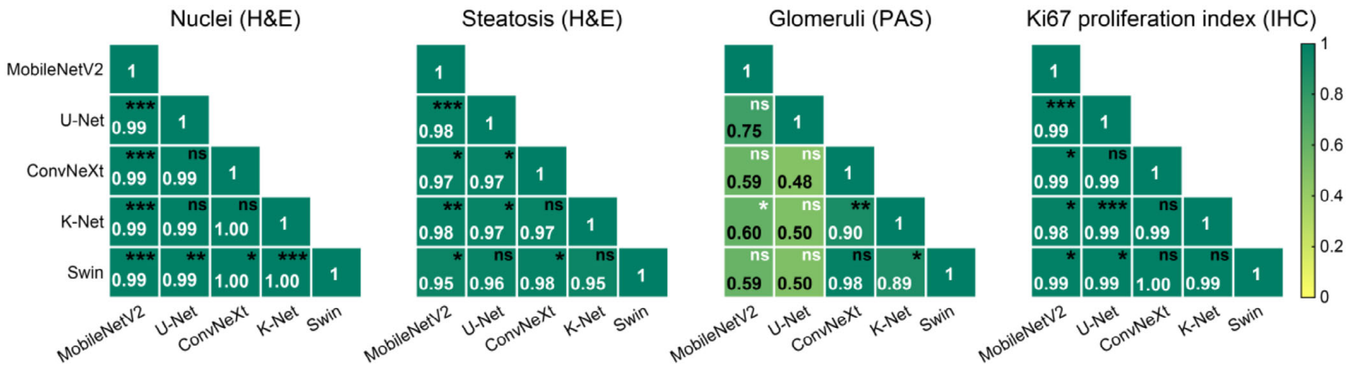

FIGURE 6 | Correlation matrix (PCC averaged over data pre-processing) with paired-sample t-tests assessing bias between quantitative results produced by any two deep learning models on WSIs for the segmentation task: nuclei count (first column), total steatosis relative area (second column), total glomeruli count (third column) and Ki67 PI (last column). Correlation matrix is by definition a square and symmetrical matrix. Legend: no statistical significance (ns); $p < 0.05$ (*); $p < 0.001$ (**); and $p < 0.0001$ (***).

For each task, we carried out two performance analyses: the PCC with paired-sample t-tests averaged over pre-processing techniques, to measure the relationships between all possible pairs of deep learning models, and the RAD metric (as defined in Equation 1) to measure the model deviation in prediction using different pre-processing techniques.

Figure 6 presents correlation matrices comparing network architecture performances across the four segmentation tasks at the WSI level. Each matrix shows the PCC averaged over all pre-processing techniques, followed by the paired-sample *t*-tests evaluating quantitative outputs yielded by two different models. The PCC indicates the strength of linear correlation between model outputs, while t-tests reveal potential systematic biases between model pairs.

For nuclei counting, all models showed extremely high correlation (PCC > 0.99), indicating strong agreement in their count predictions across WSIs. However, the t-test results revealed that only certain model pairs achieved unbiased predictions, suggesting that while the counting patterns were highly correlated, some models systematically produced higher or lower counts than others.

The Ki67 PI assessment and steatosis quantification demonstrated similarly high correlations (PCC > 0.95) across all model pairs. Notably, the more complex architectures K-Net and Swin showed no significant bias in their pairwise comparison, while most other comparisons revealed systematic differences despite strong correlations observed between the measurements.

Glomerular counts were more diverse. The lightweight models (MobileNetV2 and U-Net) exhibited lower PCC values with higher variability in these correlations compared to other architectures. In contrast, the more complex networks (ConvNeXt, K-Net, and Swin) maintained strong correlations among themselves (PCC > 0.89). Notably, only ConvNeXt and Swin achieved unbiased predictions in their pairwise comparison, suggesting these architectures might be more reliable for this specific task.

Figure 7 illustrates the RAD distribution on WSIs (mean values with standard error bars) across deep learning models for all quantitative measurements. This plot allows for the evaluation of the impact of pre-processing independently of network complexity. NORM consistently led to the lowest RAD values across most tasks: 7.14% in nuclei counting (compared to 9.65–9.82% for other pre-processing methods), 19.1% for total steatosis measurements (compared to 52.2–53.4%), and 36.7% for total glomeruli counts (compared to 58.0–49.0%). These results suggest that the NORM pre-processing step reduces the variability between the outputs of different networks, making them more consistent regardless of model complexity. The only exception was observed for the Ki67 PI assessment, where CAUG showed slightly better performance, though all pre-processing methods produced comparable results with RAD values ranging between 6.21% and 7.37%.

NORM consistently led to the lowest mean RAD values; hence, we compared the RAD distribution among the five segmentation models trained using the stain normalized data pre-processing technique (Figure 8). In terms of RAD distribution on WSIs, for the nuclei (H and E) task, ConvNeXt showed the best performance (upper whisker: 10.2%), while lightweight models (MobileNetV2 and U-Net) showed the poorest performance (maximum upper whisker: 29.5%). For Steatosis (H and E) task, all networks showed comparable distributions (maximum upper whisker: 62.8%). For the Glomeruli (PAS) task, complex models (ConvNeXt, K-Net, and Swin) outperformed others (maximum upper whisker: 117%). For the Ki67 proliferation index (IHC) task, ConvNeXt achieved the best results (upper whisker: 14.8%), while lightweight models performed worst (maximum upper whisker: 26.5%).

Finally, we performed a computational time analysis on WSIs, summarized in Figure 9. The processing times for a single WSI across different models and pre-processing techniques are provided in the supplementary material (Table S10). For the ORIG and colour augmented images, the total time includes tissue detection, network inference, and post processing steps. For stain-normalized images, an additional normalization step (30–50 s, depending on WSI dimensions) is included.

As expected, model complexity directly impacts WSI processing time. The lightweight MobileNetV2 was the fastest for all tasks and regardless of the pre-processing technique used, while the

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by Cochraneltalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

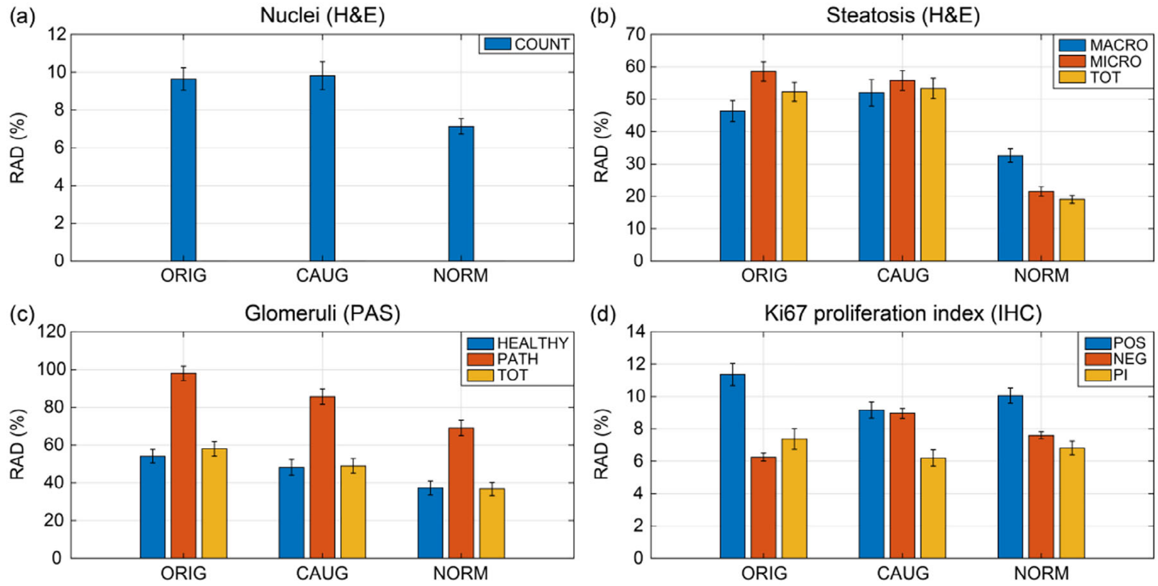

FIGURE 7 | Quantitative comparison of different data pre-processing techniques: ORIG, CAUG and NORM. Distribution of RAD (mean value with standard error bars) among various network models on WSIs: (a) nuclei count; (b) relative area of MACRO, MICRO, and total (TOT) steatosis; (c) count of healthy, pathological (PATH) and total (TOT) glomeruli; and (d) Ki67-positive (POS) and Ki67-negative (NEG) cells count and Ki67 PI.

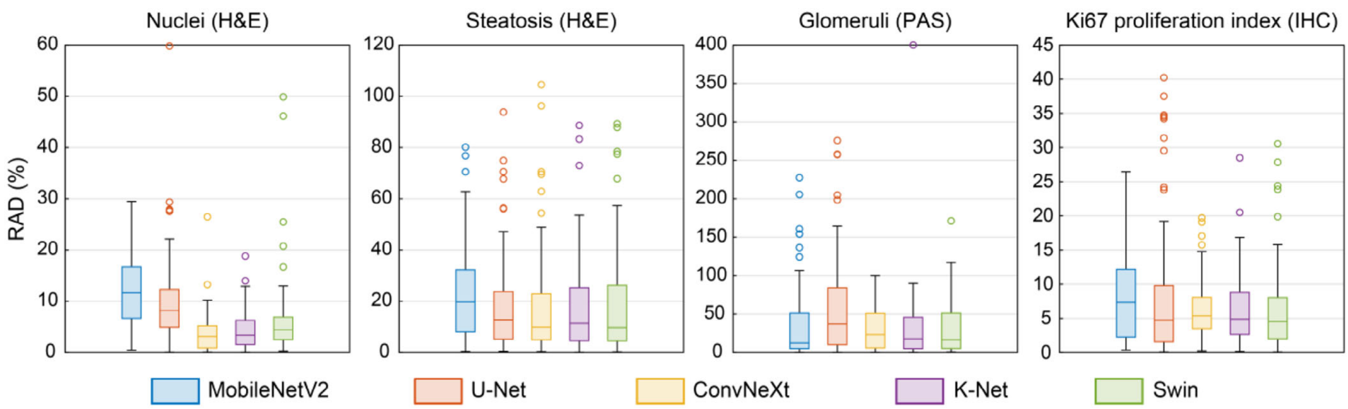

FIGURE 8 | Boxplots of the RAD metric on WSIs among various network models trained using stain normalized data pre-processing technique, for each segmentation task: nuclei count (first column), total steatosis relative area (second column), total glomeruli count (third column), and Ki67 PI (last column).

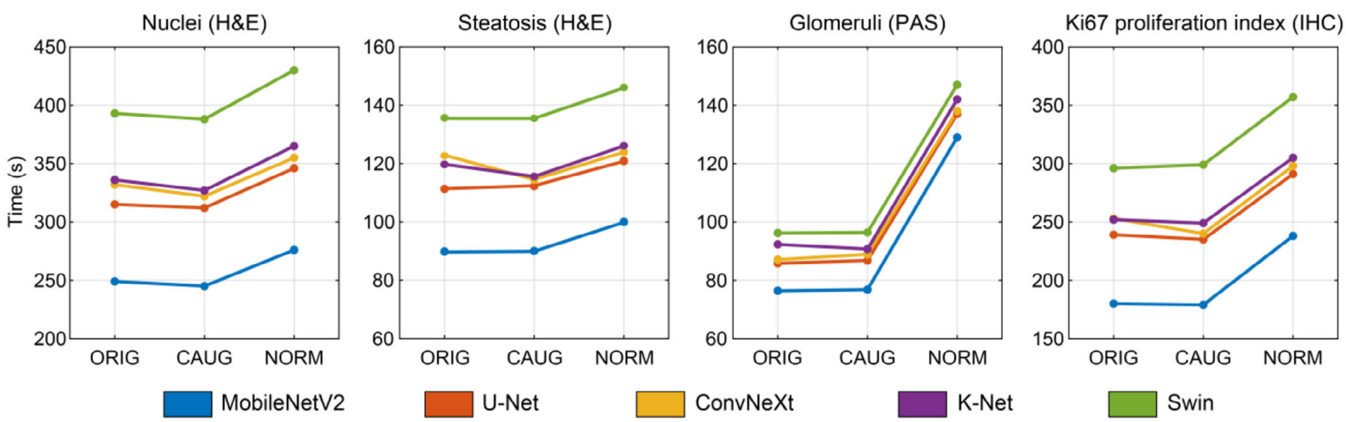

FIGURE 9 | Average WSI processing times by the five models (MobileNetV2, U-Net, ConvNeXt, K-Net, Swin) as a function of the WSI pre-processing technique (ORIG, CAUG, NORM) and analysed across four distinct digital pathology tasks.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Swin architecture consistently required more time to process WSIs. For example, to perform nuclei segmentation in WSIs, MobileNetV2 required on average 249 s/WSI, compared to Swin, which required 393 s/WSI. The NORM induced a moderate processing overhead when compared to the ORIG tile and color augmented tile processing approaches due to the additional normalization step. This is particularly evident in tasks like glomeruli detection with MobileNetV2 where processing time increased from 76.4 s (ORIG tile processing) to 129 s (processing of stain normalized WSIs). Task complexity increased computational overhead as well: segmentation of nuclei typically required longer processing times due to the large number of objects to be detected and counted in H&E WSIs. Notably, for most applications, both lightweight (MobileNetV2) and moderately complex models (U-Net, ConvNeXt, K-Net) with NORM maintained shorter computational times compared to the more complex Swin trained using either ORIG or color augmented tiles. This suggests that implementing NORM with simpler architectures can provide an efficient balance between performance and processing speed.

4 | Discussion

The in-depth examination of four digital pathology tasks questions common assumptions about the need for advanced deep learning architectures for daily segmentation applications. Our findings challenge the common assumption that more complex deep learning architectures are necessary for achieving optimal performance in computational pathology. Instead, we demonstrate that for many routine diagnostic tasks, the combination of appropriate pre-processing techniques with lightweight architectures can achieve comparable or superior results to more complex models. This represents a significant shift from current trends in the field and has important implications for the practical implementation of AI in clinical settings. Our results provide important new information about how pre-processing techniques, model complexity, and real-world performance influence one another.

4.1 | Impact of Model Complexity and Data Preprocessing

By examining the results obtained on the ORIG test image tiles without stain perturbation (Table 4)—which resemble images from a single-centre environment—we observed that higher performance (more accurate segmentation and/or counting) is not always attributed to a higher model complexity. ConvNeXt, for example, had the lowest RAE_{COUNT} (7.62%) in nuclei counting, but this error was only slightly lower when compared to U-Net (8.62%) and K-Net (8.63%). Similarly, the U-Net architecture performed better than more complex models in steatosis quantification, with an AE_{TOT} of 0.338% as opposed to 0.475% achieved by ConvNeXt.

Our findings strongly suggest that model performance is more significantly affected by data pre-processing techniques than by the model complexity alone. This is especially evident in the evaluation involving stain-perturbed tiles (Table 5), where training of the models using stain normalized tiles led to lower errors when compared to the errors of models trained using the ORIG images and/or color augmented tiles. The effect was particularly evident in the evaluation of the Ki67 PI, where the NORM decreased the AE_{PI} across all models from ranges of 8.56–19.9% (ORIG images) to 5.09–6.87% (normalized images). In addition, lightweight architectures benefit greatly from the inclusion of NORM as pre-processing. For example, MobileNetV2 trained using stain normalized tiles outperformed the same network trained on ORIG images and achieved a RAE_{COUNT} of 12.2% in nuclei counting, which was similar to Swin's performance (11.4%).

CAUG showed an intermediate position in terms of effectiveness: it provided moderate performance improvements compared to using ORIG images, but did not reach the accuracy gains achieved with NORM. For instance, CAUG decreased the AE_{PI} range in Ki67 PI assessment from 8.56–19.9% (ORIG images) to 7.32–12.8%, but it fell short of the improvement attained with stain normalization (5.09–6.87%).

The main advantage of CAUG lies in its computational efficiency during the inference phase. While NORM requires additional processing time during inference (increasing processing time by roughly 5–10% per image), CAUG adds no computational overhead during deployment since it is applied only during training. At inference time, models trained with CAUG process images without any additional computational overhead compared to models trained on ORIG images. This makes CAUG an attractive option for scenarios where processing time is critical, offering a balance between improved performance and computational efficiency.

While our general findings favour simpler architecture with proper pre-processing, our analysis reveals a more nuanced relationship between the task and model complexity. For less challenging tasks like nuclei counting and steatosis quantification, lightweight architectures with proper pre-processing achieve comparable or superior results to complex models. However, tasks involving learning of complex morphological features (such as the glomerular morphology) benefit from applying more complex models, which yield more accurate results than simpler or less complex models. This is particularly evident in glomeruli segmentation, where the target structures present intricate internal architecture including capillary tufts, Bowman's space, and various cellular components, combined with the challenge of multiclass classification (healthy vs. pathological glomeruli). This complexity in both structure and classification demands higher model capacity to effectively capture and process these sophisticated morphological features. This is evidenced by the correlation analysis (Figure 6), where complex networks (ConvNeXt, K-Net, and Swin) maintain high correlations (PCC > 0.89) and achieve superior performance compared to lightweight architectures regardless of the pre-processing technique. This task-model dependency suggests that only some, but not all digital pathology applications can benefit from a one-model-fits-all-analytical-tasks strategy.

4.2 | Practical Implementation Recommendations

The computational time analysis reveals significant differences in processing requirements between architectures. With a proper training tile pre-processing strategy, MobileNetV2 can maintain

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

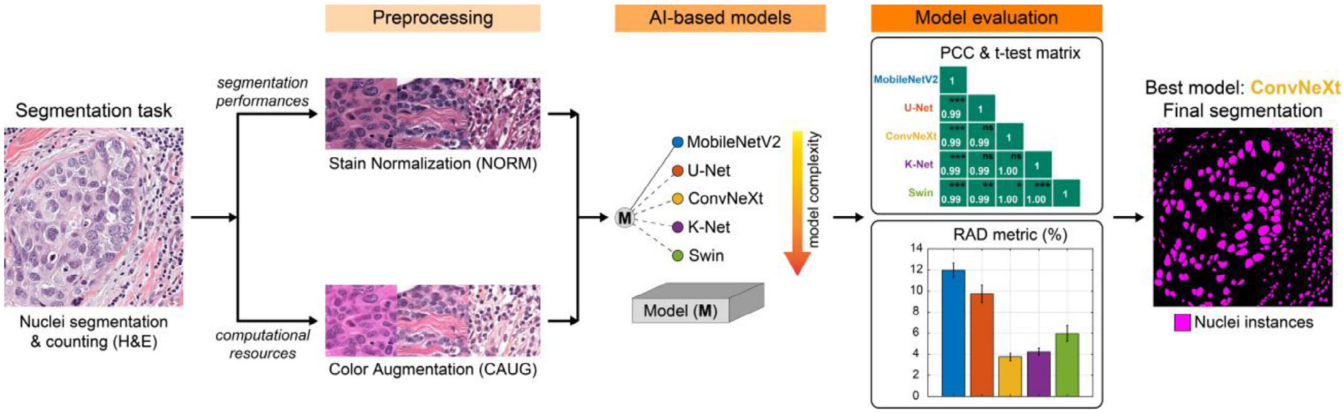

FIGURE 10 | Proposed strategy for implementing deep learning models in clinical digital pathology workflows. The approach emphasizes balancing model complexity with preprocessing techniques to achieve optimal performance while maintaining practical computational efficiency. The workflow shows the nuclei segmentation and counting task in H and E-stained images: if the aim is the optimization of segmentation performances, NORM preprocessing technique should be selected and, after model evaluation, ConvNeXt proves to be the best one, since it reaches the lowest RAD metric and shows high PCC values with no bias introduction across the five AI-based models.

high and acceptable performance levels, and process WSIs about 1.6 times faster than Swin. In light of these findings, we offer a practical strategy for integrating deep learning into digital pathology workflows:

1. Establish a robust WSI and tile pre-processing pipeline, preferably including NORM (or CAUG for applications with very limited computational resources).
2. Evaluate the performance of all candidate models using correlation metrics and statistical tests (PCC and *t*-test) along with the RAD metric to ensure performance stability.
3. Select the least complex model that demonstrates no statistically significant difference in performance compared to more complex alternatives.

This approach can help maintain balance between model accuracy and its computational efficiency, which can be particularly important in clinical settings where quick turnaround times are crucial (Figure 10). Furthermore, such optimization strategies can contribute to the democratization and widespread adoption of digital and computational pathology, making advanced analysis tools accessible to a broader range of healthcare institutions, including those with limited computational infrastructure.

Correlation coefficients and t-test results can help find the optimal balance between model performance and inference times (Figure 6). If correlation values are consistently high (PCC $>$ 0.95) across the correlation matrix and t-tests show no significant differences, this indicates that even simple models can achieve performance levels similar to those of complex models, without systematic biases. However, when a model is not suitable for a particular task, a clear correlation drop-off will emerge in the correlation matrices, where less complex models show poor correlation with more complex models, or statistical tests reveal significant differences in predictions. Such correlation degradation indicates when model simplification compromises reliability. This pattern is clearly demonstrated in our study: nuclei counting, steatosis quantification, and Ki67 assessment showed consistently high correlations (PCC $>$ 0.95) across all model pairs, indicating that lightweight models were sufficient. In contrast, for glomeruli detection, only the more complex models (ConvNeXt, K-Net, and Swin) maintained high correlations (PCC $>$ 0.89), while lightweight models showed poor correlation, suggesting that this task requires more sophisticated architectures.

Since GT annotations at the WSI level are often not available for many tasks (e.g., nuclei segmentation), we propose using the RAD as a metric to evaluate the 'similarity' of network outputs from a clinical/application perspective. This approach does not require the availability of GT, making it especially useful in situations where manual annotation of large-scale WSIs is impractical. This is exactly what happened in the glomeruli counting task (Figure 6), where lightweight networks like U-Net and MobileNetV2 exhibited weak correlation with more complex models and significant *t*-test results, suggesting that more complex networks may be necessary for achieving clinically relevant accuracy. In contrast, ConvNeXt, with intermediate complexity, showed strong correlation with the more complex models and no significant differences in the t-tests, offering a potential balance between speed and accuracy.

4.3 | Clinical Implications

There are several clinical implications that can be derived from our study. By assuming that more complex models performed better, and neglecting the relevance of pre-processing steps, pathologists (and investigators at large) may obtain suboptimal results. Multiple diagnostic histopathological criteria rely on the accurate identification (i.e., segmentation) and quantification (i.e., counting) of cells and structures (i.e., objects), both in neoplastic and non-neoplastic diseases [39–43]. Even small deviations from GT in nuclei detection, classification, and counting can significantly impact patient management, affecting diagnosis, tumor grading, and prognosis in various cancer types [44, 45]. This is particularly critical in tasks such as Ki67 PI assessment, where accuracy directly influences clinical decision-making. In

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by Cochraneltalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

such cases, models achieving optimal IoU and lowest AE_{PI} should be prioritized for clinical implementation.

Differences in object counts, as those observed in nuclei count, and in the estimation of Ki67 PI, could have significant diagnostic consequences, especially for tumour diagnosis and grading. Exemplificative of this consideration are the grading system of neuroendocrine tumors [44, 45], prognostic and predictive evaluation of breast cancers in the adjuvant and neoadjuvant settings [46], but also more practical tasks such as the evaluation of tumour tissue samples cellularity for subsequent molecular testing [47–49]. Our approach could also benefit tasks requiring precise quantitative assessment of histological patterns, which are notoriously challenging when performed manually [50].

Furthermore, our analysis of multicentric-like images suggests that, in certain contexts, pre-processing approaches may be more important than model complexity. We believe that this evidence should be highly emphasized, especially considering the crucial role of pathologists and histopathologic analysis in multicentric clinical trials development [51, 52]. Lastly, the computational time analysis of WSIs underscores the need to tailor both pre-processing methods and deep learning models to the clinical-diagnostic context, particularly from a time-sensitive perspective. Some areas of pathology require rapid, thorough evaluation of tissue samples, such as in intraoperative consultations (e.g., frozen section analysis) or donor organ assessments for transplantation [17, 53]. The additional time needed for tissue slide scanning and AI analysis could be a barrier in these scenarios. As the digital pathology field moves forward, we showed that AI-based approaches balancing improved performance with computational efficiency may be a potential solution to this technological challenge.

4.4 | Limitations and Future Directions

The main limitation of the study is that it was focused on specific segmentation tasks and established pre-processing techniques. This, however, left room to explore alternative approaches. For example, further research can include domain adaptation strategies [54], which specifically address domain shifts between centres, and self-supervised learning approaches [55], which use unlabelled data to increase model robustness.

As demonstrated in this work, the impact of preprocessing techniques cannot be predicted a priori, as it is a task and context dependent. In our experiments, NORM consistently improved performance; however, the extent of this improvement differed significantly between tasks. The complexity of multi-centre studies is evident in how performance varies by specific task. Stain variations across institutions stem from several interconnected factors: differences in laboratory procedures, reagent variability, scanner-specific characteristics, and inconsistencies in tissue preparation protocols. Different histological features and tasks may also be impacted differently by each of these factors too. Our findings suggest that pre-processing strategies should be evaluated on a task-specific basis, as their effectiveness cannot be generalized across all digital pathology applications. These findings highlight how crucial it is to conduct a systematic evaluation of pre-processing when creating solutions for clinical applications, especially in multi-centre settings.

Several important avenues for further investigation arise from these results. First, consideration should be given to the creation of lightweight, task-adaptive architectures. These designs could provide the best balance between performance and computational efficiency by dynamically adjusting their complexity in response to task requirements. For example, these networks could automatically scale their depth or width depending on the specific pathology feature being analysed. Second, research into more effective pre-processing techniques is still needed. Current NORM methods, while effective, introduce slight computational overhead (around 5–10%). Future methods could leverage recent advancements in lightweight real-time image processing frameworks and efficient neural style transfer approaches (which can adapt image styles with minimal computational overhead) to achieve similar pre-processing benefits while reducing computational cost. Third, this research field could benefit from new metrics that specifically assess clinical significance of segmentation networks. While metrics like RAE and AE provide technical performance measures, translating these into clinically meaningful thresholds remains challenging. Such benchmarks should incorporate input from pathologists regarding acceptable performance variations for different diagnostic tasks.

In this study, we did not consider the tile size. For some models, doubling the patch size does not necessarily result in a x4 increase in computational time. Future research should explore the utilization of models that are insensitive to tile size and/or can scale tiles without performance degradation—a desirable feature in analysing WSIs.

These future directions align with the broader goal of democratizing digital and computational pathology by making WSI analysis more accessible and reliable in the clinical setting. Our results strongly suggest that, for many routine digital pathology tasks, investing in robust pre-processing pipelines may be more beneficial than pursuing increasingly complex model architectures, particularly when considering the current constraints of clinical implementation.

5 | Conclusion

This study emphasizes the importance of pre-processing techniques, such as NORM, in enhancing the consistency and performance of deep learning models in digital pathology. While more complex models do not always outperform lightweight ones, NORM consistently reduced variability across all segmentation tasks. This work provides a paradigm shift from the prevailing 'bigger is better' approach towards more efficient, task-optimized solutions suitable for routine clinical implementation. This highlights the need to balance model complexity with computational efficiency, especially in time-sensitive clinical contexts. We also provide practical recommendations for the development and deployment of AI models in digital pathology. Lightweight models should be prioritized, with complexity tailored to specific task requirements. Additionally, the proposed RAD metric and PCC with t-test matrix can help ensure reliable performance across different network architectures. By focusing on pre-processing

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and efficiency, we offer a pathway to more reliable, clinically applicable AI solutions in pathology.

Author Contributions

Massimo Salvi: conceptualization, methodology, supervision, writing – original draft. **Nicola Michielli:** data curation, formal analysis, methodology, validation, visualization, writing – original draft. **Alessandro Mogetta:** data curation, methodology, visualization, writing – review and editing **Alessandro Gambella:** formal analysis, writing – review and editing. **Abdulkadir Sengur:** writing – review and editing. **Filippo Molinari:** supervision, writing – review and editing. **Arkadiusz Gertych:** validation, writing – original draft.

Acknowledgements

We would like to thank Dr. Luca Molinaro (Molinette Hospital, Turin, Italy) and Dr. Martino Bosco (Michele and Pietro Ferrero Hospital, Cuneo, Italy) for their valuable advice and support during the digitization phase of the samples.

Open access publishing facilitated by Politecnico di Torino, as part of the Wiley - CRUI-CARE agreement.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The private datasets used and analysed during the current study are available from the corresponding author on reasonable request.

References

1. V. Baxi, R. Edwards, M. Montalto, and S. Saha, "Digital Pathology and Artificial Intelligence in Translational Medicine and Clinical Practice," *Modern Pathology* 35, no. 1 (2022): 23–32, https://doi.org/10.1038/s41379-021-00919-2.

2. W. Ding, W. Liao, X. Zhu, and H. Zhu, "A Novel Automatic Annotation Method for Whole Slide Pathological Images Combined Clustering and Edge Detection Technique," *IET Image Process* 18, no. 6 (2024): 1516–1529, https://doi.org/10.1049/ipr2.13045.

3. H. Xu, Q. Xu, F. Cong, et al., "Vision Transformers for Computational Histopathology," *IEEE Reviews in Biomedical Engineering* 17 (2024): 63–79.

4. Z. Chen, T. Xie, S. Chen, et al., "AI-Based Tumor-infiltrating Lymphocytes Scoring System for Assessing HCC Prognosis in Patients Undergoing Liver Resection," *JHEP Reports* 7, no. 2 (2024): 101270.

5. K. He, J. Zhu, L. Li, F. Gou, and J. Wu, "Two-Stage Coarse-to-Fine Method for Pathological Images in Medical Decision-Making Systems," *IET Image Process* 18, no. 1 (2024): 175–193, https://doi.org/10.1049/ipr2.12941.

6. S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, A. Madabhushi, and M. Abedalthagafi, "Artificial Intelligence in Neuro-Oncology: Advances and Challenges in Brain Tumor Diagnosis, Prognosis, and Precision Treatment," *NPJ Precis Oncol* 8 (2024): 80, https://doi.org/10.1038/s41698-024-00575-0.

7. Y. Lu, J. Zhang, X. Liu, et al., "Prediction of Breast Cancer Metastasis by Deep Learning Pathology," *IET Image Process* 17, no. 2 (2023): 533–543, https://doi.org/10.1049/ipr2.12652.

8. B. Moxley-Wyles and R. Colling, "Artificial Intelligence and Digital Pathology: Where are We Now and What are the Implementation Barriers?" *Diagnostic Histopathology* 30 (2024): 597–603.

9. D. Y. Zhang, A. Venkat, H. Khasawneh, R. Sali, V. Zhang, and Z. Pei, "Implementation of Digital Pathology and Artificial Intelligence in Routine Pathology Practice," *Laboratory Investigation* 104, no. 9 (2024): 102111, https://doi.org/10.1016/j.labinv.2024.102111.

10. C. Bruce, I. Prassas, M. Mokhtar, et al., "Transforming Diagnostics: The Implementation of Digital Pathology in Clinical Laboratories," *Histopathology* 85, no. 2 (2024): 207–214, https://doi.org/10.1111/his.15178.

11. A. H. Song, G. Jaume, D. F. K. Williamson, et al., "Artificial Intelligence for Digital and Computational Pathology," *Nature Reviews Bioengineering* 1 (2023): 930–949, https://doi.org/10.1038/s44222-023-00096-8.

12. H. R. Tizhoosh and L. Pantanowitz, "Artificial Intelligence and Digital Pathology: Challenges and Opportunities," *Journal of Pathology Informatics* 9 (2018): 38, https://doi.org/10.4103/jpi.jpi_53_18.

13. I. Kim, K. Kang, Y. Song, and T.-J. Kim, "Application of Artificial Intelligence in Pathology: Trends and Challenges," *Diagnostics* 12, no. 11 (2022): 2794, https://doi.org/10.3390/diagnostics12112794.

14. L. Deininger, B. Stimpel, A. Yuce, et al. A Comparative Study Between Vision Transformers and CNNs in Digital Pathology, ArXiv Preprint ArXiv:2206.00389 (2022).

15. M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The Impact of Pre-and Post-Image Processing Techniques on Deep Learning Frameworks: A Comprehensive Review for Digital Pathology Image Analysis," *Computers in Biology and Medicine* 128 (2021): 104129, https://doi.org/10.1016/j.compbiomed.2020.104129.

16. M. Salvi, A. Mogetta, A. Gambella, et al., "Automated Assessment of Glomerulosclerosis and Tubular Atrophy Using Deep Learning," *Computerized Medical Imaging and Graphics* 90 (2021): 101930, https://doi.org/10.1016/j.compmedimag.2021.101930.

17. A. Gambella, M. Salvi, L. Molinaro, et al., "Improved Assessment of Donor Liver Steatosis Using Banff Consensus Recommendations and Deep Learning Algorithms," *Journal of Hepatology* 80, no. 3 (2024): 495–504, https://doi.org/10.1016/j.jhep.2023.11.013.

18. J. Gallego, Z. Swiderska-Chadaj, T. Markiewicz, M. Yamashita, M. A. Gabaldon, and A. Gertych, "A U-Net Based Framework to Quantify Glomerulosclerosis in Digitized PAS and H and E Stained Human Tissues," *Computerized Medical Imaging and Graphics* 89 (2021): 101865, https://doi.org/10.1016/j.compmedimag.2021.101865.

19. M. Salvi, A. Mogetta, U. Raghavendra, A. Gudigar, U. R. Acharya, and F. Molinari, "A Dynamic Uncertainty-Aware Ensemble Model: Application to Lung Cancer Segmentation in Digital Pathology," *Applied Soft Computing* 165 (2024): 112081, https://doi.org/10.1016/j.asoc.2024.112081.

20. M. Cui and D. Y. Zhang, "Artificial Intelligence and Computational Pathology," *Laboratory Investigation* 101, no. 4 (2021): 412–422, https://doi.org/10.1038/s41374-020-00514-0.

21. N. Kumar, R. Verma, D. Anand, et al., "A Multi-Organ Nucleus Segmentation Challenge," *IEEE Transactions on Medical Imaging* 39, no. 5 (2019): 1380–1391, https://doi.org/10.1109/TMI.2019.2947628.

22. M. Salvi, L. Molinaro, J. Metovic, et al., "Fully Automated Quantitative Assessment of Hepatic Steatosis in Liver Transplants," *Computers in Biology and Medicine* 123 (2020): 103836, https://doi.org/10.1016/j.compbiomed.2020.103836.

23. D. Tellez, G. Litjens, P. Bándi, et al., "Quantifying the Effects of Data Augmentation and Stain Color Normalization in Convolutional Neural Networks for Computational Pathology," *Medical Image Analysis* 58 (2019): 101544, https://doi.org/10.1016/j.media.2019.101544.

24. N. Altini, G. D. Cascarano, A. Brunetti, et al., "Semantic Segmentation Framework for Glomeruli Detection and Classification in Kidney Histological Sections," *Electronics (Basel)* 9, no. 3 (2020): 503.

25. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018): 4510–4520.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by CochraneItalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

26. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (2015): 234–241.

27. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A Convnet for the 2020s,” in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9 (2022): 11976–11986.

28. W. Zhang, J. Pang, K. Chen, and C. C. Loy, “K-Net: Towards Unified Image Segmentation,” *NIPS’21: Proceedings of the 35th International Conference on Neural Information Processing Systems* 34 (2021): 10326–10338.

29. Z. Liu, Y. Lin, Y. Cao, et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021): 10012–10022.

30. M. Salvi, N. Michielli, and F. Molinari, “Stain Color Adaptive Normalization (SCAN) Algorithm: Separation and Standardization of Histological Stains in Digital Pathology,” *Computer Methods and Programs in Biomedicine* 193 (2020): 105506, https://doi.org/10.1016/j.cmpb.2020.105506.

31. N. Michielli, A. Caputo, M. Scotto, et al., “Stain Normalization in Digital Pathology: Clinical Multi-Center Evaluation of Image Quality,” *Journal of Pathology Informatics* 13 (2022): 100145, https://doi.org/10.1016/j.jpi.2022.100145.

32. A. Gertych, N. Zurek, N. Piaseczna, et al., “Tumor Cellularity Assessment Using AI Trained on IHC-Restained Slides Improves Selection of Lung Adenocarcinoma Samples for Molecular Testing,” *American Journal of Pathology* 195, no. 5 (2025): 907–922, https://doi.org/10.1016/j.ajpath.2025.01.009.

33. A. Basu, P. Senapati, M. Deb, R. Rai, and K. G. Dhal, “A Survey on Recent Trends in Deep Learning for Nucleus Segmentation From Histopathology Images,” *Evolving Systems* 15 (2024): 203–248, https://doi.org/10.1007/s12530-023-09491-3.

34. G. E. Liquori, G. Calamita, D. Cascella, M. Mastrodonato, P. Portincasa, and D. Ferri, “An Innovative Methodology for the Automated Morphometric and Quantitative Estimation of Liver Steatosis,” *Histology and Histopathology* 24, no. 1 (2009): 49–60.

35. V. G. Puelles, W. E. Hoy, M. D. Hughson, B. Diouf, R. N. Douglas-Denton, and J. F. Bertram, “Glomerular Number and Size Variability and Risk for Kidney Disease,” *Current Opinion in Nephrology and Hypertension* 20, no. 1 (2011): 7–15, https://doi.org/10.1097/MNH.0b013e3283410a7d.

36. K. Kontzoglou, V. Palla, G. Karaolanis, et al., “Correlation Between Ki67 and Breast Cancer Prognosis,” *Oncology* 84, no. 4 (2013): 219–225, https://doi.org/10.1159/000346475.

37. N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, “A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology,” *IEEE Transactions on Medical Imaging* 36 (2017): 1550–1560, https://doi.org/10.1109/TMI.2017.2677499.

38. Mms. Contributors, MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark, (2020).

39. R. Salgado, C. Denkert, S. Demaria, et al., “The Evaluation of Tumor-Infiltrating Lymphocytes (TILs) in Breast Cancer: Recommendations by an International TILs Working Group 2014,” *Annals of Oncology* 26, no. 2 (2015): 259–271, https://doi.org/10.1093/annonc/mdu450.

40. A. Lugli, R. Kirsch, Y. Ajioka, et al., “Recommendations for Reporting Tumor Budding in Colorectal Cancer Based on the International Tumor Budding Consensus Conference (ITBCC) 2016,” *Modern Pathology* 30, no. 9 (2017): 1299–1311, https://doi.org/10.1038/modpathol.2017.46.

41. A. W. Lohse, M. Sebode, P. S. Bhathal, et al., “Consensus Recommendations for Histological Criteria of Autoimmune Hepatitis From the International AIH Pathology Group: Results of a Workshop on AIH Histology Hosted by the European Reference Network on Hepatological Diseases and the European Society of Pathology,” *Liver International* 42, no. 5 (2022): 1058–1069.

42. M. Campora, M. Paudice, A. Gambella, et al., “Counting Mitoses in Gastrointestinal Stromal Tumours (GISTs): Variable Practices in the Real-World Setting and Their Clinical Implications,” *Virchows Archiv* 482 (2023): 589–594, https://doi.org/10.1007/s00428-022-03454-w.

43. V. Deshpande, Y. Zen, J. K. C. Chan, et al., “Consensus Statement on the Pathology of IgG4-Related Disease,” *Modern Pathology* 25, no. 9 (2012): 1181–1192, https://doi.org/10.1038/modpathol.2012.72.

44. C. Luchini, L. Pantanowitz, V. Adsay, et al., “Ki-67 Assessment of Pancreatic Neuroendocrine Neoplasms: Systematic Review and Meta-Analysis of Manual vs. digital Pathology Scoring,” *Modern Pathology* 35, no. 6 (2022): 712–720, https://doi.org/10.1038/s41379-022-01055-1.

45. S. La Rosa, “Diagnostic, Prognostic, and Predictive Role of Ki67 Proliferative Index in Neuroendocrine and Endocrine Neoplasms: Past, Present, and Future,” *Endocrine Pathology* 34 (2023): 79–97, https://doi.org/10.1007/s12022-023-09755-3.

46. A. H. Skjervold, H. S. Pettersen, M. Valla, S. Opdahl, and A. M. Bofin, “Visual and Digital Assessment of Ki-67 in Breast Cancer Tissue-a Comparison of Methods,” *Diagnostic Pathology* 17 (2022): 45, https://doi.org/10.1186/s13000-022-01225-4.

47. H. H. Popper, J. Tímár, A. Ryska, and W. Olszewski, “Minimal Requirements for the Molecular Testing of Lung Cancer,” *Translational Lung Cancer Research* 3, no. 5 (2014): 301.

48. D. A. Eberhard, G. Giaccone, and B. E. Johnson, “Biomarkers of Response to Epidermal Growth Factor Receptor Inhibitors in Non–Small-Cell Lung Cancer Working Group: Standardization for Use in the Clinical Trial Setting,” *Journal of Clinical Oncology* 26 (2008): 983–994, https://doi.org/10.1200/JCO.2007.12.9858.

49. E. Thunnissen, K. M. Kerr, F. J. F. Herth, et al., “The Challenge of NSCLC Diagnosis and Predictive Analysis on Small Samples. Practical Approach of a Working Group,” *Lung Cancer* 76, no. 1 (2012): 1–18, https://doi.org/10.1016/j.lungcan.2011.10.017.

50. A. Gertych, Z. Swiderska-Chadaj, Z. Ma, et al., “Convolutional Neural Networks Can Accurately Distinguish Four Histologic Growth Patterns of Lung Adenocarcinoma in Digital Slides,” *Scientific Reports* 9 (2019): 1483, https://doi.org/10.1038/s41598-018-37638-9.

51. X. Matias-Guiu, M. R. Raspollini, J. Kulka, et al., “The Role of the Pathologist in the Design and Conducting of Biomarker-driven Clinical Trials in Cancer: Position Paper of the European Society of Pathology,” *Virchows Archiv* (2024): 1–8.

52. C. Röcken, “Quality Assurance in Clinical Trials—The Role of Pathology,” *Virchows Archiv* 468 (2016): 83–92.

53. M. Salvi, A. Mogetta, K. M. Meiburger, et al., “Karpinski Score Under Digital Investigation: A Fully Automated Segmentation Algorithm to Identify Vascular and Stromal Injury of Donors’ kidneys,” *Electronics* 9, no. 10 (2020): 1644.

54. R. Deng, C. Cui, Q. Liu, et al., Segment Anything Model (Sam) for Digital Pathology: Assess Zero-Shot Segmentation on Whole Slide Imaging, ArXiv Preprint ArXiv:2304.04155 (2023).

55. H. Wang, E. Ahn, and J. Kim, “A Multi-resolution Self-Supervised Learning Framework for Semantic Segmentation in Histopathology,” *Pattern Recognition* 155 (2024): 110621, https://doi.org/10.1016/j.patcog.2024.110621.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting Information file 1: ipr270290-sup-0001-SuppMat.docx.

17519667, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70290 by Cochraneitalia, Wiley Online Library on [27/01/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License