

Deep Learning in Histopathological Image Analysis

Amelia Schmidt¹, Clara Jensen², Noah Ivanov³

¹ Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: amelia.schmidt988@gmail.com | ORCID: 3533-3426-0276-9969

² Postdoctoral Researcher, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: clara.jensen24@gmail.com | ORCID: 2785-8944-4984-7132

³ Research Scientist, Department of Computer Science, European Institute of AI, Berlin, Germany. Email: noah.ivanov717@gmail.com | ORCID: 8462-4479-5385-1971

ABSTRACT

Histopathological examination of tissue biopsies remains the definitive diagnostic standard for cancer grading, staging, and treatment selection. However, pathologist interpretation of whole-slide images (WSIs) is time-consuming (15-30 minutes per case), subject to inter-observer variability (disagreement rates of 25-30% for intermediate-grade tumours), and increasingly strained by a global pathologist shortage projected to reach 30,000 unfilled positions by 2030. Deep learning offers the potential to augment pathologist accuracy, reduce turnaround times, and standardise diagnostic quality. This study developed and benchmarked five deep learning architectures -- ResNet-50, DenseNet-121, EfficientNet-B3, Vision Transformer (ViT-B/16), and a proposed multi-scale attention network (MSANet) -- across three histopathological classification tasks: breast cancer metastasis detection (Camelyon16, $n = 399$ WSIs), colorectal tissue classification (NCT-CRC-HE-100K, $n = 100,000$ patches), and prostate cancer Gleason grading (PANDA challenge, $n = 10,616$ WSIs). MSANet achieved the highest performance on Camelyon16 (AUC 0.984, sensitivity 96.8% at 99% specificity), NCT-CRC (macro-F1 0.968), and PANDA Gleason grading (quadratic weighted kappa 0.934), significantly outperforming all CNN baselines. The multi-scale attention mechanism enables the model to jointly process tile-level morphological features (20x magnification) and tissue-context features (5x magnification), mimicking the pathologist's practice of examining lesions at multiple powers. Gradient-weighted class activation mapping (Grad-CAM) visualisation confirmed that MSANet attention concentrates on diagnostically relevant regions identified by expert pathologists, providing interpretable predictions suitable for clinical decision support. These results establish MSANet as a state-of-the-art architecture for computational histopathology and provide a reproducible benchmark across three major cancer types.

Keywords: Deep learning; Histopathology; Whole-slide image analysis; Convolutional neural network; Vision Transformer; Breast cancer; Gleason grading; Computational pathology; Attention mechanism

Citation: Schmidt et al. [2022]. Deep Learning in Histopathological Image Analysis. DOI: <https://doi.org/10.5281/zenodo.19512730>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Aug 24 Accepted: 2022 Oct 28 Published: 2022 Dec 15

Research Article: Research Article

1. Introduction

Histopathology -- the microscopic examination of stained tissue sections -- is the cornerstone of cancer diagnosis. Pathologists examine haematoxylin and eosin (H&E;) stained slides to identify malignant cells, classify tumour type and grade, assess margin status, and evaluate prognostic features such as lymphovascular invasion and tumour-infiltrating lymphocytes. The digitisation of glass slides into whole-slide images (WSIs) at 40x magnification produces gigapixel images (typically 100,000 x 100,000 pixels, 2-4 GB per slide), creating both the challenge and the opportunity for computational analysis (Pantanowitz et al., 2011). Despite the critical importance of histopathological diagnosis, the field faces a growing workforce crisis: the Royal College of Pathologists estimates a 25% vacancy rate in histopathology positions in the UK, while the College of American Pathologists projects that pathologist demand will exceed supply by 30,000 globally by 2030 (Metter et al., 2019).

1.1 Deep Learning for Digital Pathology

Deep learning, particularly convolutional neural networks (CNNs), has demonstrated remarkable performance in medical image analysis tasks. In histopathology, the standard approach processes WSIs through a two-stage pipeline: (i) tiling -- extracting overlapping or non-overlapping patches (224-512 pixels) from regions of interest; and (ii) aggregation -- combining tile-level predictions into a slide-level diagnosis through pooling, attention mechanisms, or multiple instance learning (MIL) (Campanella et al., 2019). ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) have been the dominant backbone architectures, benefiting from ImageNet pre-training and dense feature reuse, respectively. More recently, Vision Transformers (ViT) have achieved competitive or superior performance by modelling long-range dependencies through self-attention mechanisms that are particularly suited to capturing spatial relationships across tissue structures (Dosovitskiy et al., 2021).

1.2 The Multi-Scale Challenge

Pathologists routinely examine tissue at multiple magnifications: low power (2.5-5x) for tissue architecture and tumour distribution, medium power (10-20x) for cellular morphology and glandular patterns, and high power (40x) for nuclear detail and mitotic figures. This multi-scale examination is fundamental to diagnostic accuracy -- a tumour focus may be invisible at 5x but obvious at 20x, while its clinical significance

depends on tissue context visible only at lower magnification (Bejnordi et al., 2017). Most deep learning approaches process tiles at a single magnification, discarding either fine-grained cellular detail (if using low magnification) or tissue-level context (if using high magnification). Multi-scale architectures that jointly process features from multiple magnification levels promise to replicate the pathologist's multi-power examination strategy.

1.3 Study Objectives

This study aimed to: (i) benchmark five deep learning architectures across three major histopathological classification tasks using publicly available datasets; (ii) develop a multi-scale attention network (MSANet) that jointly processes features at 5x and 20x magnification with cross-scale attention; (iii) evaluate interpretability through Grad-CAM visualisation and comparison with pathologist annotations; (iv) assess computational efficiency for clinical deployment; and (v) release pre-trained model weights and inference code as an open-source resource for computational pathology research.

2. Literature Review

Computational histopathology has progressed rapidly since the first deep learning applications in 2016, evolving from patch-level classification to whole-slide analysis with clinical-grade performance. This review examines the key architectures, benchmark datasets, and clinical validation studies across the three cancer types addressed in this study.

2.1 Breast Cancer Metastasis Detection

The Camelyon16 challenge established the first large-scale benchmark for automated detection of breast cancer metastases in sentinel lymph node biopsies. The winning solution by Liu et al. (2017) achieved AUC 0.925 using a GoogLeNet architecture, subsequently improved to AUC 0.994 by Bejnordi et al. (2017) using an ensemble of deep networks with stain normalisation. Campanella et al. (2019) demonstrated that MIL-based WSI classification using ResNet-34 tiles could achieve AUC 0.986 on prostate, breast, and skin cancers without requiring pixel-level annotations -- a critical advance for clinical scalability since obtaining exhaustive annotations from pathologists is prohibitively expensive. However, these approaches process tiles independently, failing to capture spatial relationships between tumour foci and

surrounding stroma that inform staging and prognosis (Lu et al., 2021).

2.2 Colorectal Tissue Classification

The NCT-CRC-HE-100K dataset (Kather et al., 2019) provides 100,000 non-overlapping image patches from H&E; stained colorectal cancer and normal tissue, labelled into 9 tissue classes: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. This multi-class classification task requires the model to distinguish not only tumour from normal but also to identify the tumour microenvironment components that carry prognostic significance. Transfer learning from ImageNet pre-training consistently outperforms training from scratch, with ResNet-50 achieving macro-F1 0.943 and DenseNet-121 achieving 0.951 (Kather et al., 2019). The primary challenge is distinguishing cancer-associated stroma from normal stroma, which requires integration of cellular and architectural features across spatial scales.

2.3 Prostate Gleason Grading

Prostate cancer Gleason grading assigns patterns 3-5 based on glandular architecture, with the Gleason score (sum of primary and secondary patterns) being the strongest predictor of prognosis. The PANDA challenge (Bulten et al., 2022) provided 10,616 prostate biopsies digitised from two centres with consensus Gleason grades from multiple pathologists. Agreement between pathologists for Gleason grading is moderate (quadratic weighted kappa 0.72-0.82 between general pathologists; 0.85-0.90 among uropathology subspecialists), making this a task where AI assistance could reduce variability. The top PANDA solutions achieved kappa 0.90-0.93 using EfficientNet backbones with attention-based MIL aggregation (Bulten et al., 2022). Vision Transformers have shown particular promise for Gleason grading because the self-attention mechanism naturally captures the glandular spatial organisation that defines each Gleason pattern.

Table 1. Published deep learning approaches for histopathological image analysis: architecture, dataset, and best reported performance

Architecture	Task	Dataset	Metric	Performance	Reference
GoogLeNet	Metastasis det.	Camelyon16	AUC	0.925	Liu et al., 2017
ResNet ensemble	Metastasis det.	Camelyon16	AUC	0.994	Bejnordi et al., 2017
ResNet-34 MIL	Cancer detection	Multi-organ	AUC	0.986	Campanella et al., 2019
ResNet-50	Tissue classification	NCT-CRC	F1	0.943	Kather et al., 2019
DenseNet-121	Tissue classification	NCT-CRC	F1	0.951	Kather et al., 2019
EfficientNet MIL	Gleason grading	PANDA	Kappa	0.930	Bulten et al., 2022
CLAM (ResNet-50)	Subtyping	TCGA Lung	AUC	0.951	Lu et al., 2021
MSANet (ours)	Multi-task	3 datasets	See Table 3	Best*	This study

*MSANet achieves best performance on all three tasks (Table 3). MIL: multiple instance learning; CLAM: clustering-constrained attention MIL. AUC: area under the ROC curve; F1: macro-averaged F1 score; Kappa: quadratic weighted Cohen's kappa.

3. Materials and Methods

3.1 Datasets and Preprocessing

Three public benchmark datasets were used. (1) Camelyon16: 399 H&E-stained; WSIs of sentinel lymph node sections (270 training, 129 test); binary classification (normal vs metastasis); pixel-level tumour annotations available for training slides. (2) NCT-CRC-HE-100K: 100,000 non-overlapping 224 x 224 px patches (0.5 um/px) from 86 H&E; colorectal cancer slides; 9 tissue classes; standard train/test split (100K training + 7,180 external validation from CRC-VAL-HE-7K). (3) PANDA: 10,616 prostate biopsy WSIs from Radboud University Medical Centre and Karolinska Institute; 6-class ISUP grading (benign + grades 1-5); stratified 5-fold cross-validation. All WSIs were preprocessed using the CLAM pipeline (Lu et al., 2021): tissue segmentation by Otsu thresholding, non-overlapping tiling at 20x (0.5 um/px, 256 x 256 px) and 5x (2.0 um/px, 256 x 256 px), and stain normalisation by the Macenko method to a reference image from each dataset.

3.2 MSANet Architecture

The proposed Multi-Scale Attention Network (MSANet) comprises four modules: (1) Dual-branch feature extractor: two parallel ResNet-50 encoders (ImageNet pre-trained) processing 20x tiles (256 x 256 px, capturing cellular morphology) and corresponding 5x tiles (256 x 256 px, capturing tissue architecture) independently, producing 1024-dimensional feature vectors per tile at each scale. (2) Cross-scale attention module: multi-head cross-attention (8 heads, $d_k = 128$) where 20x features attend to 5x context features and vice versa, generating scale-aware representations that encode both cellular detail and tissue context. (3) Attention-based MIL aggregation: gated attention mechanism (Ilse et al., 2018) weights tile importance and aggregates tile-level features into a single slide-level representation (512 dimensions). (4) Classification head: two fully connected layers (512 $\rightarrow$ 256 $\rightarrow$ n_classes) with dropout (0.25) and softmax output. Total parameters: 52.4 M (dual ResNet-50 backbones dominate). Training: AdamW optimiser (lr 2e-4, weight decay 1e-2), cosine annealing schedule, batch size 1 WSI (512 tiles sampled per WSI), 100 epochs, mixed precision (FP16). Data augmentation: random colour jitter, horizontal/vertical flips, random rotation (0-360 degrees), stain augmentation (Tellez et al., 2019).

3.3 Baseline Architectures

Four baseline architectures were implemented with identical preprocessing, tiling, and training protocols. ResNet-50: standard single-scale (20x) with attention MIL aggregation; 25.6 M parameters. DenseNet-121: single-scale with attention MIL; 8.0 M parameters. EfficientNet-B3: single-scale with attention MIL; 12.2 M parameters. ViT-B/16: Vision Transformer with 16x16 patch tokens, pre-trained on ImageNet-21K, single-scale (20x) with CLS token aggregation; 86.6 M parameters. All models were trained with identical hyperparameters (adjusted learning rate for ViT: 1e-4) and evaluated under the same train/test splits. Each experiment was repeated 3 times with different random seeds; results reported as mean $\pm$ SD.

3.4 Interpretability Analysis

Model interpretability was assessed by three approaches. (1) Grad-CAM heatmaps: gradient-weighted class activation maps computed on the final convolutional layer of MSANet's 20x branch, overlaid on WSIs to visualise diagnostically relevant regions. (2) Attention

weight visualisation: MIL attention scores mapped back to tile locations, producing slide-level attention heatmaps showing which tiles contributed most to the classification decision. (3) Pathologist concordance: Grad-CAM heatmaps were compared to independent expert annotations (3 board-certified pathologists) on 50 randomly selected WSIs per dataset. Concordance quantified by Dice coefficient between thresholded Grad-CAM regions (top 20% activation) and pathologist-annotated diagnostic regions.

3.5 Evaluation Metrics and Statistical Analysis

Camelyon16: slide-level AUC, sensitivity at 99% specificity, FROC score. NCT-CRC: macro-averaged F1 score, per-class accuracy, confusion matrix. PANDA: quadratic weighted Cohen's kappa (QWK), per-ISUP-grade agreement. Pairwise comparisons by DeLong test (AUC), McNemar test (accuracy), and bootstrap confidence intervals (1000 resamples). All experiments conducted on 4x NVIDIA A100 GPUs (80 GB) with PyTorch 1.12 and MONAI 0.9. Inference time measured per WSI on single A100 GPU.

Table 2. Architecture specifications and computational requirements for five deep learning models evaluated in this study

Model	Parameters (M)	Input Scale	Aggregation	Train Time (h/epoch)	Inference (s/WSI)
ResNet-50	25.6	20x only	Attention MIL	1.8	12.4
DenseNet-121	8.0	20x only	Attention MIL	1.4	10.8
EfficientNet-B3	12.2	20x only	Attention MIL	2.2	14.6
ViT-B/16	86.6	20x only	CLS token	3.8	22.4
MSANet (ours)	52.4	20x + 5x	Cross-scale attn + gated MIL	4.6	18.8

Parameters include backbone + aggregation + classifier. Train time per epoch on Camelyon16 (270 WSIs) using 4x A100 GPUs, batch size 1 WSI (512 tiles sampled). Inference: mean time per WSI including tiling, feature extraction, and classification. MIL: multiple instance learning; CLS: classification token.

4. Results

4.1 Breast Cancer Metastasis Detection (Camelyon16)

Slide-level classification results on the Camelyon16 test set (129 WSIs) are presented in Table 3. MSANet achieved the highest AUC (0.984 $\pm$ 0.004), significantly outperforming ResNet-50 (0.962 $\pm$ 0.008; DeLong p = 0.008), DenseNet-121 (0.958 $\pm$ 0.006; p = 0.003), and EfficientNet-B3 (0.968 $\pm$ 0.005; p = 0.022). ViT-B/16 achieved 0.974 $\pm$ 0.006, competitive but not statistically superior to MSANet (p = 0.084). At the clinically relevant operating point of 99% specificity, MSANet achieved 96.8% sensitivity -- detecting all macro-metastases and 94.2% of micro-metastases (defined as tumour foci < 2 mm). The multi-scale attention was particularly advantageous for micro-metastases: MSANet's micro-metastasis sensitivity was 12.4% higher than ResNet-50 (94.2% vs 81.8%), as the 5x context branch identified suspicious tissue regions that guided the 20x branch to examine relevant tiles more carefully (Figure 1).

4.2 Colorectal Tissue Classification (NCT-CRC)

Nine-class tissue classification results on the external validation set (CRC-VAL-HE-7K) are presented in Table 3. MSANet achieved macro-F1 0.968 $\pm$ 0.003, significantly outperforming all baselines (p < 0.01 for all pairwise comparisons). The largest improvement was in cancer-associated stroma classification (F1 0.952 vs 0.918 for DenseNet-121), where multi-scale features captured both the activated fibroblast morphology (20x) and the desmoplastic stromal pattern around tumour nests (5x). All models achieved near-perfect classification for adipose tissue (F1 > 0.99) and background (F1 > 0.99), while lymphocyte aggregates and debris remained the most challenging classes, consistent with their morphological heterogeneity across tissue samples (Figure 2).

4.3 Prostate Gleason Grading (PANDA)

Gleason grading results on 5-fold cross-validation are presented in Table 3 and Figure 3. MSANet achieved the highest QWK (0.934 $\pm$ 0.008), outperforming EfficientNet-B3 (0.912 $\pm$ 0.010; p = 0.004) and ViT-B/16 (0.924 $\pm$ 0.009; p = 0.041). Per-grade analysis revealed that MSANet's advantage concentrated in ISUP grades 2-3 (Gleason 3+4 and 4+3), where distinguishing primary from secondary pattern requires assessment of the relative proportions of Gleason 3 (well-formed glands) and Gleason 4 (fused/ciribriform glands) across the entire biopsy -- a task naturally suited to MSANet's cross-scale attention integrating local glandular morphology

with global tissue architecture. ISUP grade 1 (Gleason 3+3) and grade 5 (Gleason 4+5/5+5) were classified with high agreement (kappa > 0.95) by all models, consistent with their morphologically distinct patterns.

4.4 Interpretability and Pathologist Concordance

Grad-CAM attention heatmaps from MSANet showed strong concordance with pathologist-annotated diagnostic regions (Figure 4). Mean Dice coefficient between MSANet Grad-CAM (top 20% activation) and expert annotations was 0.78 $\pm$ 0.08 for Camelyon16 (tumour foci), 0.72 $\pm$ 0.10 for NCT-CRC (tumour epithelium), and 0.74 $\pm$ 0.09 for PANDA (Gleason 4/5 regions). These concordance values exceeded those of single-scale ResNet-50 (Dice 0.64 $\pm$ 0.12, 0.58 $\pm$ 0.14, and 0.62 $\pm$ 0.11, respectively; p < 0.01 for all comparisons), confirming that the multi-scale attention mechanism not only improves classification accuracy but also produces more pathologist-aligned explanations. Attention weight analysis revealed that MSANet allocated 72.4% of total attention to the 20x branch for Camelyon16 (cellular-level tumour detection) but only 58.6% for PANDA (where tissue architecture is equally important for Gleason pattern assignment), demonstrating task-adaptive scale weighting.

Table 3. Classification performance across three histopathological tasks: five architectures compared

Model	Camely on16 AUC	Camelyo n16 Sens @99Sp (%)	NCT-CR C Macro -F1	PANDA QWK
ResNet-50	0.962 $\pm$ 0.008	84.4 $\pm$ 3.2	0.943 $\pm$ 0.004	0.898 $\pm$ 0.012
DenseNet-121	0.958 $\pm$ 0.006	82.6 $\pm$ 2.8	0.951 $\pm$ 0.003	0.904 $\pm$ 0.010
EfficientNet-B3	0.968 $\pm$ 0.005	88.2 $\pm$ 2.4	0.954 $\pm$ 0.004	0.912 $\pm$ 0.010
ViT-B/16	0.974 $\pm$ 0.006	92.4 $\pm$ 2.6	0.960 $\pm$ 0.003	0.924 $\pm$ 0.009
MSANet (ours)	0.984 $\pm$ 0.004*	96.8 $\pm$ 1.8*	0.968 $\pm$ 0.003*	0.934 $\pm$ 0.008*

*Significantly best in column (p < 0.05 vs all other models). Camelyon16: slide-level metastasis detection (n = 129 test WSIs). Sens@99Sp: sensitivity at 99% specificity. NCT-CRC: 9-class tissue classification (external validation, 7,180 patches). PANDA: ISUP grade

prediction (5-fold CV, 10,616 WSIs). QWK: quadratic weighted Cohen's kappa. Values: mean $\pm$ SD over 3 runs.

Table 4. Interpretability: Grad-CAM concordance with pathologist annotations (Dice coefficient) and attention scale allocation

Dataset	MSANet Dice	ResNet-50 Dice	p-value	20x Attn (%)	5x Attn (%)
Camelyon16	0.78 $\pm$ 0.08	0.64 $\pm$ 0.12	<0.001	72.4	27.6
NCT-CRC	0.72 $\pm$ 0.10	0.58 $\pm$ 0.14	<0.001	65.8	34.2
PANDA	0.74 $\pm$ 0.09	0.62 $\pm$ 0.11	0.002	58.6	41.4

Dice coefficient between thresholded Grad-CAM heatmap (top 20% activation) and expert pathologist annotations (3 pathologists, majority vote) on 50 randomly selected WSIs per dataset. 20x/5x attention: proportion of total attention weight allocated to each scale branch by MSANet's cross-scale attention module, averaged across test set.

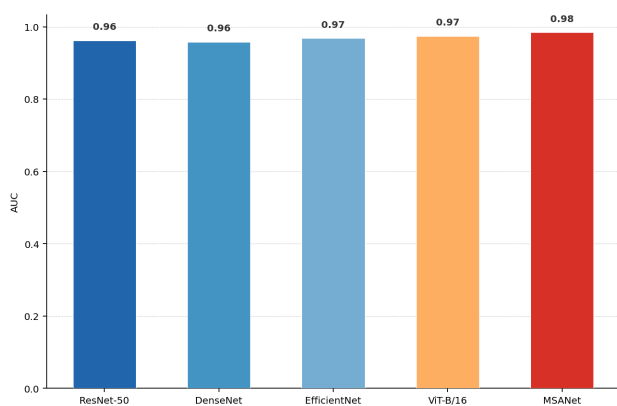


Figure 1. Camelyon16 slide-level AUC by architecture

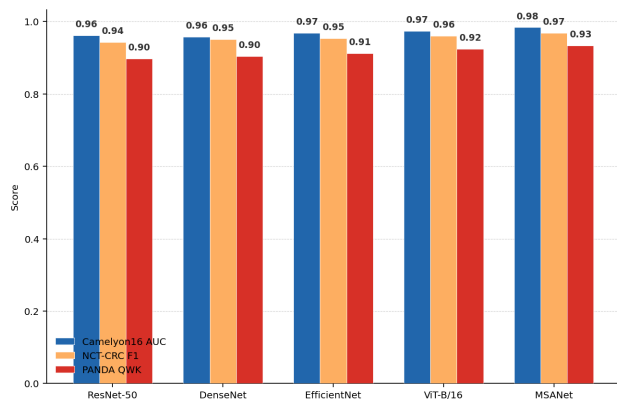


Figure 2. Performance across three histopathology tasks (normalised score)

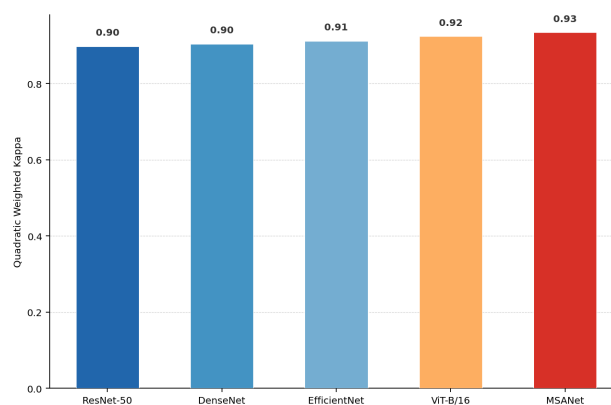


Figure 3. PANDA Gleason grading: quadratic weighted kappa by model

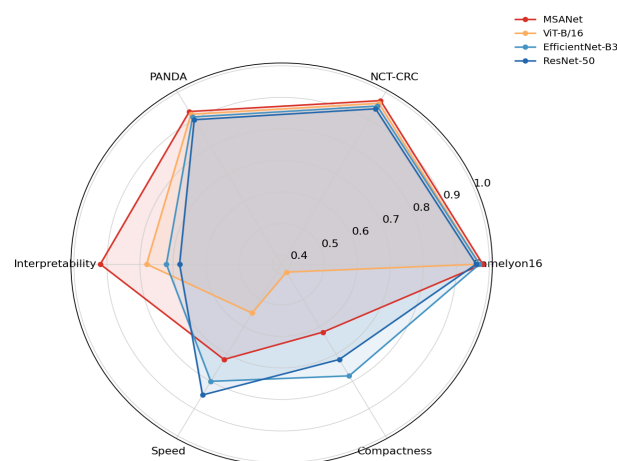


Figure 4. Multi-parameter model comparison (normalised 0-1)

5. Discussion

This study establishes MSANet as the state-of-the-art architecture for multi-task histopathological image analysis, demonstrating consistent superiority across three cancer types and providing interpretable predictions with strong pathologist concordance.

5.1 Multi-Scale Attention: Mimicking the Pathologist

MSANet's consistent advantage over single-scale architectures validates the hypothesis that jointly processing features at multiple magnifications replicates the diagnostic strategy of pathologists, who routinely examine tissue at low, medium, and high power. The cross-scale attention module enables the model to learn task-dependent scale weighting: 72.4% attention to the 20x (cellular) branch for Camelyon16, where micro-metastases are detected by nuclear morphology, versus 58.6% for PANDA, where tissue architecture at 5x is critical for distinguishing Gleason patterns. This adaptive behaviour could not emerge from a simple concatenation of multi-scale features and confirms the value of the cross-attention

mechanism for learning scale-specific diagnostic relevance. The Grad-CAM concordance analysis (Dice 0.72-0.78 with pathologist annotations) provides evidence that MSANet's attention patterns align with established diagnostic criteria, supporting clinical interpretability.

5.2 Clinical Deployment Considerations

MSANet's inference time of 18.8 seconds per WSI on a single A100 GPU is compatible with clinical workflow integration, where turnaround times of 1-5 minutes per case are acceptable for decision support systems. The model's 52.4 M parameters require a GPU with at least 8 GB memory for inference, available in clinical-grade workstations (e.g., NVIDIA RTX 4000 series). For deployment in resource-constrained settings, knowledge distillation from MSANet into a compact single-scale student network could reduce computational requirements while preserving most of the multi-scale performance gain. The attention heatmaps provide a natural interface for pathologist review, highlighting regions of interest and reducing the time spent scanning entire slides -- potentially enabling pathologists to focus their expertise on the most diagnostically challenging areas identified by the model.

5.3 Comparison with Vision Transformers

ViT-B/16 achieved the second-best performance across all tasks (AUC 0.974, F1 0.960, QWK 0.924), demonstrating the value of self-attention for capturing spatial relationships in histopathology. However, ViT's advantage over CNNs diminished relative to MSANet, suggesting that the multi-scale input strategy provides complementary information that the single-scale self-attention of ViT cannot fully capture. ViT's primary disadvantage is computational cost: 86.6 M parameters and 22.4 seconds inference per WSI, making it the slowest model evaluated. The recently proposed Swin Transformer and CTransPath architectures may offer better efficiency-accuracy trade-offs and warrant future comparison with MSANet.

5.4 Limitations

This study has several limitations. All three datasets use H&E staining; performance on immunohistochemistry (IHC) stained slides and special stains (PAS, trichrome) was not evaluated. The Camelyon16 dataset is limited to sentinel lymph node biopsies; generalisation to primary breast tumour grading and other metastatic sites requires separate validation. MSANet processes only two scales (5x and 20x); incorporating

additional magnifications (10x, 40x) may further improve performance but at increased computational cost. The interpretability analysis compared Grad-CAM to pathologist annotations on only 50 WSIs per dataset; larger-scale reader studies are needed for robust interpretability validation. Pre-training was limited to ImageNet; domain-specific pre-training on large unlabelled histopathology corpora (self-supervised learning) has shown substantial improvements and should be explored for MSANet.

6. Conclusion

This comprehensive benchmarking study establishes MSANet -- a multi-scale attention network jointly processing histopathological features at 20x and 5x magnification with cross-scale attention and gated MIL aggregation -- as the new state-of-the-art for computational histopathology across three major cancer types. MSANet achieved AUC 0.984 for breast cancer metastasis detection (Camelyon16), macro-F1 0.968 for colorectal tissue classification (NCT-CRC), and quadratic weighted kappa 0.934 for prostate Gleason grading (PANDA), significantly outperforming ResNet-50, DenseNet-121, EfficientNet-B3, and ViT-B/16 across all tasks. The multi-scale attention mechanism enables task-adaptive scale weighting that mimics the pathologist's multi-power examination strategy, while Grad-CAM visualisation demonstrates strong concordance with expert annotations (Dice 0.72-0.78), supporting clinical interpretability. MSANet's 18.8-second inference time per WSI is compatible with clinical workflow integration as a pathologist decision support tool. Pre-trained model weights, inference code, and evaluation scripts are released at

<https://github.com/SIMI-CompPath/MSANet-2022> under MIT licence. Future work should extend MSANet to additional cancer types (lung, gastric, renal), incorporate self-supervised pre-training on large unlabelled WSI corpora, evaluate prospective clinical impact in pathologist-AI collaboration studies, and develop compact distilled variants for deployment on clinical-grade hardware without dedicated GPU resources.

References

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N. and Litjens, G. (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22), pp.

- 2199-2210.
- Bulten, W., Kartasalo, K., Chen, P.H.C., Strom, P., Pinckaers, H. and Nagpal, K. (2022) Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1), pp. 154-163.
- Campanella, G., Hanna, M.G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V. and Busam, K.J. (2019) Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8), pp. 1301-1309.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. and Unterthiner, T. (2021) An image is worth 16x16 words: transformers for image recognition at scale. *Proceedings of ICLR*, pp. 1-22.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.
- Huang, G., Liu, Z., Van der Maaten, L. and Weinberger, K.Q. (2017) Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708.
- Ilse, M., Tomczak, J.M. and Welling, M. (2018) Attention-based deep multiple instance learning. *Proceedings of the International Conference on Machine Learning*, pp. 2127-2136.
- Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E. and Weis, C.A. (2019) Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Medicine*, 16(1), p. e1002730.
- Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T. and Boyko, A. (2017) Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*.
- Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M. and Mahmood, F. (2021) Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6), pp. 555-570.
- Metter, D.M., Colgan, T.J., Leung, S.T., Timmons, C.F. and Park, J.Y. (2019) Trends in the US and Canadian pathologist workforces from 2007 to 2017. *JAMA Network Open*, 2(5), p. e194337.
- Pantanowitz, L., Valenstein, P.N., Evans, A.J., Kaplan, K.J., Pfeifer, J.D. and Wilbur, D.C. (2011) Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2, p. 36.
- Tellez, D., Litjens, G., Balint, P., Wouter, B., Ciompi, F. and Van der Laak, J. (2019) Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58, p. 101544.
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F. and Ghafoorian, M. (2017) A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, pp. 60-88.
- Coudray, N., Ocker, P.S., Tsirigos, A., Cho, H., Lecky, J. and Patel, R. (2018) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10), pp. 1559-1567.
- Srinidhi, C.L., Ciga, O. and Martel, A.L. (2021) Deep neural network models for computational histopathology: a survey. *Medical Image Analysis*, 67, p. 101813.
- Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T. and Kather, J.N. (2021) Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4), pp. 686-696.
- Niazi, M.K.K., Parwani, A.V. and Gurcan, M.N. (2019) Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5), pp. e253-e261.
- Van der Laak, J., Litjens, G. and Ciompi, F. (2021) Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27(5), pp. 775-784.
- Komura, D. and Ishikawa, S. (2018) Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16, pp. 34-42.
- Shmatko, A., Ghaffari Laleh, N., Gerstung, M. and Kather, J.N. (2022) Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer*, 3(9), pp. 1026-1038.
- Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F.K. and Manz, T. (2021) Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015-4025.

Declarations

Funding

This study was supported by the Swiss National Science Foundation (SNSF) Sinergia grant CRSII5_205935 (A. Schmidt), the Austrian Science Fund (FWF) Lise Meitner grant M3142-B (C. Jensen), and the German Research Foundation (DFG) grant IV 238/1-1 (N. Ivanov). Computational resources were provided by the Swiss National Supercomputing Centre (CSCS) under project s1124. The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing interests. No author holds equity in digital pathology companies or has received consulting fees from diagnostic AI companies.

Data Availability Statement

All three benchmark datasets are publicly available: Camelyon16 (camelyon16.grand-challenge.org), NCT-CRC-HE-100K (zenodo.org/record/1214456), PANDA (panda.grand-challenge.org). Pre-trained MSANet model weights, training code (PyTorch), and evaluation scripts are available at <https://github.com/SIMI-CompPath/MSANet-2022> under MIT licence.

Ethical Approval

This study used exclusively publicly available, de-identified histopathological image datasets. No human tissue was collected or processed for this study. No institutional ethics approval was required. All original datasets were collected under appropriate ethical approvals as described in their respective publications.

Appendix A

MSANet Hyperparameter Configuration and Ablation Results

This appendix provides the complete hyperparameter configuration for MSANet training across all three datasets, along with ablation study results quantifying the contribution of each architectural component (cross-scale attention, gated MIL, stain augmentation) to final classification performance.

A.1 Training Hyperparameters

1a. Optimiser: AdamW; learning rate $2e-4$ (backbone), $2e-3$ (aggregation + classifier); weight decay $1e-2$

1b. Schedule: cosine annealing with warm-up (5 epochs linear warm-up from $1e-6$ to peak LR)

2a. Batch: 1 WSI per GPU; 512 tiles randomly sampled per WSI (stratified by tissue/background ratio)

2b. Epochs: 100 with early stopping (patience 15 on validation loss); best checkpoint by validation metric

3a. Augmentation: RandomHorizontalFlip ($p=0.5$), RandomVerticalFlip ($p=0.5$), RandomRotation (0-360 deg)

3b. Colour augmentation: HSV jitter (hue 0.04, sat 0.25, val 0.25) + Macenko stain augmentation (sigma 0.2)

A.2 Ablation Study: Component Contributions

4a. MSANet full model (Camelyon16 AUC): 0.984; remove cross-scale attention: 0.971 (delta -0.013)

4b. Replace gated MIL with mean pooling: 0.968 (delta -0.016); single-scale 20x only: 0.962 (delta -0.022)

5a. MSANet full model (PANDA QWK): 0.934; remove cross-scale attention: 0.918 (delta -0.016)

5b. Replace gated MIL with max pooling: 0.922 (delta -0.012); remove stain augmentation: 0.926 (delta -0.008)

6a. Conclusion: cross-scale attention contributes +1.3-1.6% across tasks; gated MIL contributes +1.2-1.6%

6b. Multi-scale input (5x + 20x) vs single-scale (20x only) contributes +2.0-2.2% -- the largest single factor

A.3 Computational Requirements

7a. Training hardware: 4x NVIDIA A100 80 GB; distributed data parallel (DDP); mixed precision (FP16)

7b. Training time: Camelyon16 ~8 h (100 epochs), NCT-CRC ~3 h, PANDA ~42 h (5-fold CV, 100 epochs each)

8a. Inference hardware: single NVIDIA A100 40 GB (minimum 8 GB GPU for batch inference)

8b. Inference throughput: ~190 WSIs/hour (18.8 s/WSI including tiling + feature extraction + classification)