

Efficiency and Interpretability in MYC Status Prediction: A Comparative Study of Vision Transformers

GEI KI TANG

geiki@studentmail.unimap.edu.my

Universiti Malaysia Perlis

CHEE CHIN LIM

Universiti Malaysia Perlis

FAEZAHTUL ARBAEYAH HUSSAIN

Universiti Sains Malaysia

QI WEI OUNG

Universiti Malaysia Perlis

AIDY IRMAN YAJID

Universiti Sains Malaysia

SUMAYYAH MOHAMMAD AZMI

Universiti Sains Malaysia

Article

Keywords:

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9834504/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on September 21st, 2026. See the published version at <https://doi.org/10.1038/s41598-026-70959-8>.

A Comparative Study of Vision Transformer Architectures Evaluating MYC Status Prediction in Diffuse Large B Cell Lymphoma

Gei Ki Tang^{1,*}, Chee Chin Lim^{1,2,*}, Faezahtul Arbaeyah Hussain^{3,4,+}, Qi Wei Oung^{1,5,+}, Aidy Irman Yajid^{3,4,+}, and Sumayyah Mohammad Azmi^{3,4,+}

¹Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis, Arau, Perlis, 02600, Malaysia

²Sport Engineering Research Centre, Universiti Malaysia Perlis, Arau, Perlis, 02600, Malaysia

³Department of Pathology, School of Medical Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, Kelantan, 16150, Malaysia

⁴Hospital Pakar Universiti Sains Malaysia, Kubang Kerian, Kelantan, 16150, Malaysia

⁵Advanced Communication Engineering (ACE) Centre of Excellence, Universiti Malaysia Perlis, Kangar, Perlis, 01000, Malaysia

*geiki@studentmail.unimap.edu.my; cclim@unimap.edu.my

+these authors contributed equally to this work

ABSTRACT

Predicting MYC proto-oncogene (MYC) status in Diffuse Large B-Cell Lymphoma (DLBCL) is essential for patient treatment planning. However, the gold-standard laboratory test is Fluorescence in Situ Hybridization (FISH). This test is expensive and takes a long time to deliver results. In order to assist or pre-screen for the FISH test, this study evaluates deep learning models. We establish a comprehensive benchmark comparing five distinct Vision Transformer (ViT) architectures. These include Basic ViT, T2T-ViT, CLIP-ViT, Swin Transformer, and Hierarchical ViT. The primary input consists of digital Hematoxylin and Eosin (H&E) whole-slide images (WSIs). We extracted and normalized 463,200 high-resolution tissue patches from 120 slides to output a binary prediction of MYC-positive or MYC-negative status. We evaluated the models on predictive accuracy, Area Under the Receiver Operating Characteristic (AUROC), and computational inference throughput measured in iterations per second (it/s). The results reveal clear trade-offs for digital pathology workflows. The Hierarchical ViT achieved the highest absolute classification accuracy because it excels at analyzing multi-scale biological features. Meanwhile, DeLong's test showed that the Swin Transformer maintained a small but statistically significant edge in risk stratification (AUROC = 0.90 vs. 0.89, $p = 0.040$). The Swin Transformer also delivered nearly three times faster inference throughput at 5.16 it/s compared to 1.78 it/s for the Hierarchical ViT. Qualitative explainability maps using Gradient-weighted Class Activation Mapping (Grad-CAM) showed that these high-performing models focus on morphologically relevant tumor regions. This study provides a structural blueprint under single-centre and patch-level constraints. It proves that the Swin Transformer optimizes processing speed while the Hierarchical ViT maximizes diagnostic accuracy.

Introduction

Diffuse large B-cell lymphoma (DLBCL) represents the most common form of non-Hodgkin lymphoma globally, characterized by significant genetic and clinical heterogeneity¹. Among the various molecular drivers, rearrangements of the *MYC* oncogene are of paramount clinical importance; *MYC*-positive (*MYC*+) DLBCL is associated with aggressive disease progression, frequent central nervous system involvement, and poor response to standard chemoimmunotherapy². Consequently, the rapid identification of *MYC* status is essential for risk stratification and the selection of intensified therapeutic regimens³.

Current diagnostic protocols, primarily Fluorescence In Situ Hybridization (FISH) and immunohistochemistry (IHC), are resource-intensive and prone to inter-observer variability^{4,5}. While digital pathology has enabled deep learning-based biomarker prediction from Hematoxylin and Eosin (H&E) slides, the computational challenge lies in the multi-scale structural complexity of histopathological images. Unlike natural images, lymphoma tissue is characterized by complex, multi-scale morphological features. These features range from subtle nuclear pleomorphism at the microscopic level to global cellular density across the slide⁶.

Historically, Convolutional Neural Networks (CNNs) have been the standard for medical imaging. However, their inherent "locality" bias often fails to capture the long-range spatial dependencies necessary for understanding global tissue context. Vision Transformers (ViTs) have emerged as a powerful alternative, utilizing self-attention mechanisms to model these global

relationships. Yet, a "one-size-fits-all" transformer approach is insufficient for the hierarchical nature of oncology.

Incorporating Explainable AI (XAI) within digital pathology is essential for bridging the gap between deep learning metrics and clinical utility. Because standard deep learning networks function as black boxes, visual interpretations like attention heatmaps are vital for building clinical trust and validating that models are learning actual, physically relevant biological features rather than background artifacts⁷.

Individual vision transformers have been sporadically adapted for computational pathology. However, a systematic evaluation contrasting their core structural mechanisms within hematopathology remains entirely absent. To the best of our knowledge, this study represents the first comprehensive, head-to-head benchmark evaluating five structurally diverse transformer paradigms simultaneously for the prediction of *MYC* translocation status in DLBCL whole-slide images (WSIs). Rather than merely deploying standard architectures, this selection was systematically curated to map how fundamentally different architectural inductive biases handle the multi-scale cellular and structural complexities of lymphoma tissue:

1. Basic ViT: To evaluate the efficacy of pure global self-attention.
2. Tokens-to-Token (T2T) ViT: To assess whether recursive token aggregation better captures local structural information.
3. CLIP-ViT B/16: To investigate the transferability of features from large-scale multi-modal pre-training.
4. Swin Transformer: To evaluate the impact of shifted-windowing and localized attention.
5. Hierarchical ViT: To mimic the pathologist's multi-resolution workflow through a pyramid-like feature extraction.

By analysing these architectures against a large-scale dataset of 463,200 patches, we aim to identify the optimal balance between computational efficiency (throughput) and predictive interpretability, providing a robust framework for automated biomarker screening in DLBCL workflows.

Related Work

The evolution of automated diagnostics for DLBCL has undergone a rapid paradigm shift, transitioning from localized feature extraction to complex, multi-scale spatial modelling.

CNNs historically established the baseline for deep learning in lymphoma. ResNet-based architectures could categorize lymphoma subtypes^{8,9}. However, these models are intrinsically limited by their "locality" bias¹⁰. While CNNs excel at capturing local cellular features, they often struggle to model the long-range spatial dependencies required to identify genetic rearrangements like *MYC* status, which manifest as heterogeneous tissue alterations. Recent comparative reviews⁶ emphasize that CNNs frequently fail to maintain performance when faced with the high-resolution global context of gigapixel WSIs.

To address the limitations of CNNs, the field has pivoted toward ViTs, which utilize self-attention mechanisms to model interactions across non-adjacent image regions. In 2024, specialized architectures were introduced of specialized architectures like HistoTransformer and Slide-ViT, designed specifically to handle the unique computational overhead of gigapixel pathology slides¹¹. These models began to demonstrate that capturing global tissue context is essential for predicting complex biomarkers. Simultaneously, researchers explored the transferability of high-capacity feature representations from "Foundation Models" (e.g., CLIP-based architectures) to bridge the domain gap between general computer vision and specialized oncology without exhaustive domain-specific retraining.

As of 2026, the digital pathology frontier has further expanded to include domain-specific foundation models such as CONCH¹², Virchow¹³, and Prov-GigaPath¹⁴, which leverage self-supervised learning on billions of tissue patches. While these models offer state-of-the-art representations, our study deliberately benchmarks standard, accessible transformer paradigms (such as Swin and Hierarchical ViTs) to establish baseline structural trade-offs when operating within computational constraints common to typical institutional laboratories¹⁵. This focus provides a valuable baseline, as deploying multi-billion parameter foundation models often introduce severe infrastructure, deployment, and regulatory complexities that standard end-to-end architectures can bypass in localized environments.

Concurrently, the research frontier has moved toward optimizing the Efficiency-Interpretability trade-off¹⁶. Lymphoma tissue contains diagnostic hallmarks at multiple scales, from the subcellular level of the nucleus to the architectural level of tissue density¹⁷. Consequently, hierarchical models are more biologically appropriate than "flat" transformer blocks for this type of analysis, as their localized attention windows and multi-resolution feature extraction are well suited for modelling both minute cellular pleomorphism and global tissue context simultaneously¹⁸. While accuracy in these models has reached high levels, their relative performance and computational throughput for *MYC* status in DLBCL have remained under-explored^{19,20}.

Despite these advancements, a significant gap persists in the literature. There is a lack of rigorous, head-to-head benchmarking that simultaneously evaluates predictive accuracy, inference throughput (it/s), and biological explainability (XAI) across distinct transformer paradigms. Our study addresses this void by evaluating five architectures (Basic, T2T, CLIP-based, Swin, and Hierarchical) to identify the "Pareto-optimal" solution. Hence, our study aims to identify a model that provides the highest diagnostic reliability for *MYC* prediction while maintaining the computational efficiency required for modern digital pathology workflows.

Materials and Methods

Dataset and Image Acquisition

The study used a cohort of 120 high-resolution WSIs of DLBCL tissue sections stained with H&E. The cohort was balanced, consisting of 60 cases confirmed as *MYC*⁺ and 60 cases as *MYC*⁻. These images were obtained from the Department of Pathology, Hospital Universiti Sains Malaysia. Image acquisition was executed using a Motic EasyScan Pro digital slide scanner with 0.25 $\mu\text{m}/\text{pixel}$, which is the standard for 40 $\times$ Motic scanning at 40 $\times$ magnification, mapping the 120 digital slides to exactly 120 unique patients (one slide per patient) to ensure diagnostic independence across the balanced cohorts. This research was carried out in full compliance with the ethical principles outlined in the Declaration of Helsinki. The study protocol underwent rigorous independent review and was granted formal approval by the Human Research Ethics Committee of Universiti Sains Malaysia (Jawatankuasa Etika Penyelidikan Manusia Universiti Sains Malaysia; JEPeM-USM) on April 2, 2023. The investigation was conducted under the authorized protocol reference: USM/JEPeM/22110749.

The dataset partition was strictly enforced at the patient and WSI level to ensure rigorous evaluation and avoid any structural data leakage. Patches originating from a single WSI or patient were confined entirely to a single partition. The cohort of 120 WSIs (from 120 distinct patients) was split into a Training set ($N = 84$ WSIs; 324,240 patches), a Validation set ($N = 12$ WSIs; 46,320 patches), and a Testing set ($N = 12$ WSIs; 46,320 patches), maintaining a strict 80:10:10 distribution across both *MYC*⁺ and *MYC*⁻ classes. This large-scale sampling strategy was designed to provide the transformer architectures with sufficient morphological diversity to capture subtle oncogenic signatures. Table 1 tabulates the dataset distribution of this study.

Table 1. Dataset Distribution and Ethical Summary

Class	<i>MYC</i> ⁺	<i>MYC</i> ⁻
Number of WSIs	60	60
Number of Patches	231,600	231,600
Ethics Ref: USM/JEPeM/22110749		

Data Pre-processing and Computational Pipeline

The data pre-processing workflow was implemented as a structured, sequential pipeline designed to transform high-resolution WSIs into a format optimized for transformer-based feature extraction while meticulously preserving critical oncogenic signals. The initial phase involved the application of lossless compression to the raw WSIs. Given the massive scale of digital pathology files, this step was essential to balance storage efficiency with the preservation of high-frequency morphological data. Unlike lossy alternatives, lossless compression ensures that fine-grained cytological details remain undistorted. These details, such as nuclear chromatin texture and nucleolar prominence, are primary indicators of *MYC* rearrangement. This preservation of high-fidelity information is critical for the subsequent attention mechanisms of the ViTs to accurately capture subtle biological signatures²¹. Figure 1 illustrates the workflow diagram of this study.

Following compression, the images underwent Macenko normalization for ensuring color constancy across the dataset. Digital histopathology is frequently susceptible to technical variations in staining intensity and slide preparation, a phenomenon that may introduce artifactual biases into the models²². Using the Macenko method, the H&E-stained slides were decomposed into their respective optical density components and remapped to a standardized reference template. This standardization process mitigates the risk of the models learning laboratory-specific artifacts, thereby forcing the architectures to prioritize relevant histological features such as cellular density and nuclear-to-cytoplasmic ratios^{22,23}.

While Macenko normalization effectively unifies stain profiles across standard datasets, its mathematical framework relies on the presence of typical H&E optical density distributions. Consequently, the method exhibits distinct limitations when processing severely degraded or atypically stained DLBCL slides. In specimens exhibiting extensive tumor necrosis, crush artifacts, or hyper-chromatic nuclear over-staining, conditions common in high-grade lymphoma biopsies, the algorithm's singular value decomposition (SVD) matrix can fail to accurately decouple the baseline stain vectors. This failure can result in artificial color shifting, localized pixel clipping, or the unintended suppression of hyper-chromatic structural signatures critical for diagnosing cellular atypia. Acknowledging these edge cases is vital, as heavily compromised slides may necessitate robust deep-learning-based color augmentation frameworks or manual quality control gating before model inference. Figure 2 shows examples of normalization achieved through the Macenko method.

The final stage of the pre-processing pipeline consisted of a high-density patch extraction strategy: Each normalized WSI was systematically tiled into non-overlapping patches, each measuring 256 by 256 pixels²⁴. To maximize the diagnostic relevance of the training data, an automated filtration process was employed to exclude background regions and low-cellularity artifacts²⁵. The number of excluded patches was slide-dependent as WSIs vary inherently in tissue dimensions. Hence, an automated low-entropy filtration pipeline was implemented to standardize this process objectively. Patches were evaluated based

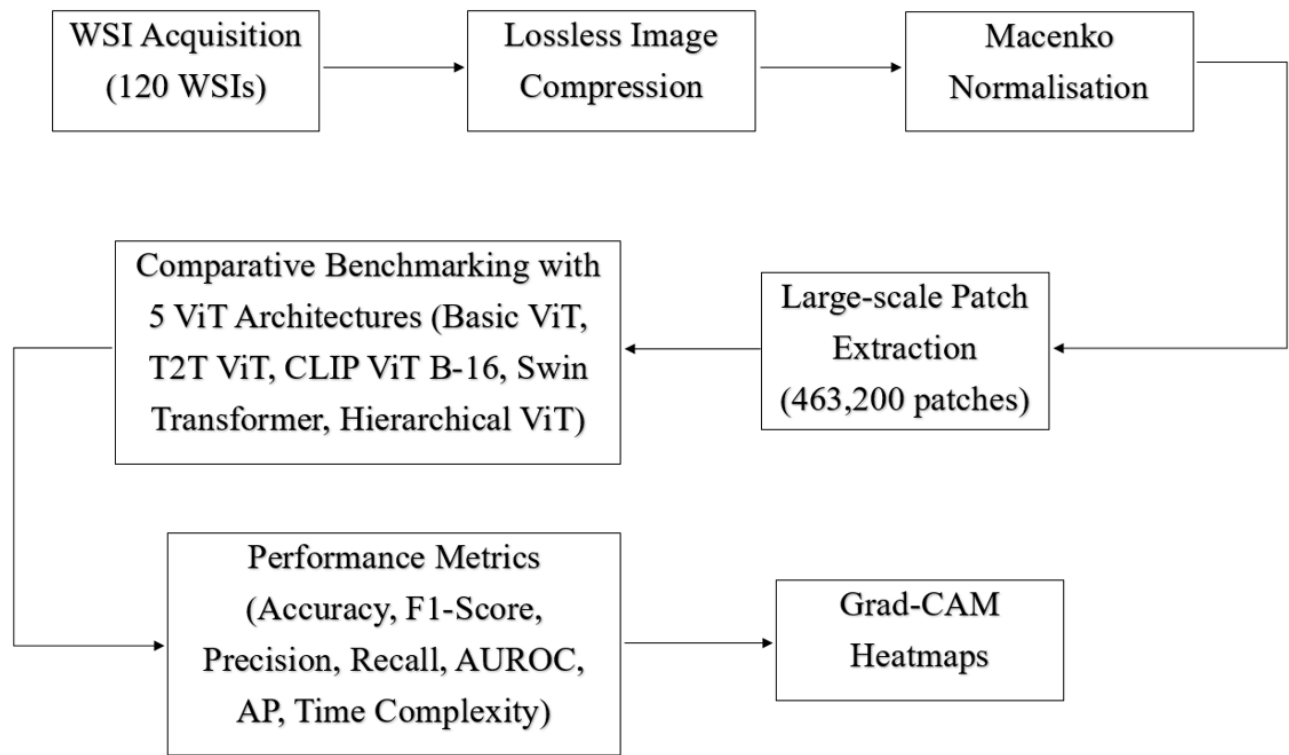


Figure 1. Workflow pipeline of this study

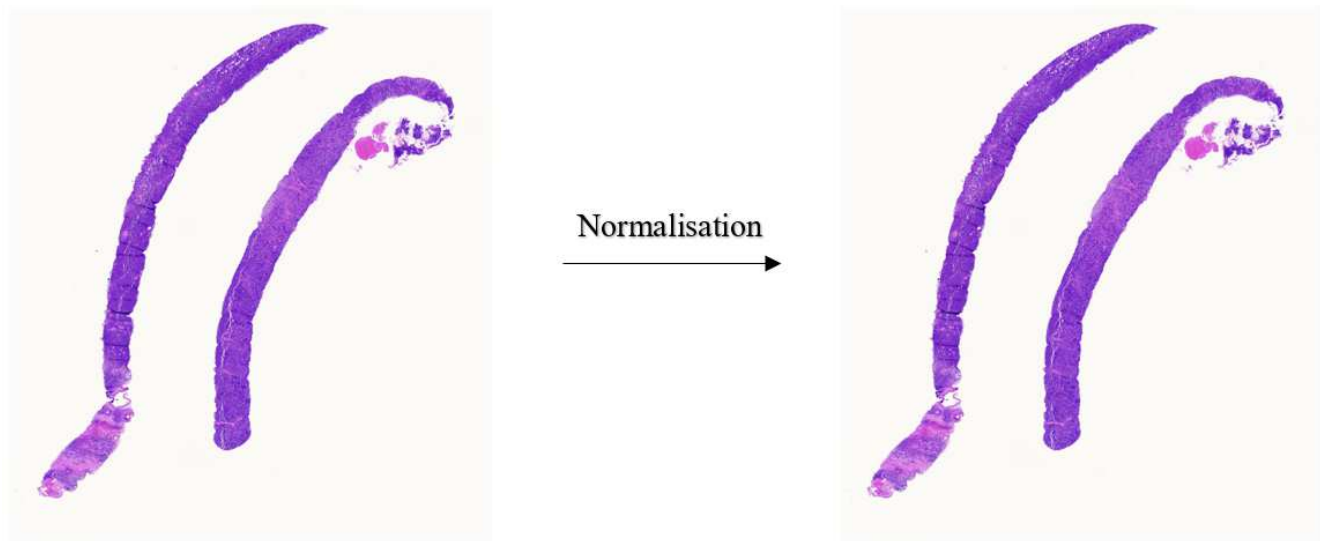


Figure 2. Normalization Samples by using Macenko Normalisation Method. (a) shows original WSI and (b) Normalised WSI

on their compressed data density. Those falling below a strict spatial complexity threshold, where the structural information content aligned with uninformative background or tissue-free whitespace, were dynamically pruned. This high-throughput preprocessing step isolated uninformative zones while preserving 100% of high-density diagnostic regions, yielding a finalized baseline of 463,200 patches.

Consequently, this process yielded a comprehensive dataset comprising 463,200 patches, meticulously balanced with 231,600 patches allocated to each class (*MYC+* and *MYC-*). These patches subsequently constituted the input for the comparative benchmarking of the five ViT architectures. The entire pipeline commences with WSI acquisition and concludes with the generation of Grad-CAM heatmaps. This pipeline establishes a robust and reproducible framework for evaluating the predictive capacity of hierarchical versus global attention mechanisms. Ultimately, it serves to enhance biomarker detection in the context of DLBCL.

Vision Transformers Architectures

The selection of these five ViT architectures was strategically engineered to compare how varying attention mechanisms handle the multi-scale morphology of DLBCL. While clinical biomarkers are ultimately adjudicated at the patient level, this study deliberately benchmarks performance at the patch level ($N = 463,200$). This granular approach isolates the direct feature-extraction capabilities of the transformer architectures. It effectively maps how distinct structural inductive biases interpret localized tumor heterogeneity, sub-visual microenvironments, and fine-grained cellular pleomorphism without the mathematical smoothing or compounding errors often introduced by slide-level aggregation layers.

Specifically, the generic classification heads were replaced with task-specific Multi-Layer Perceptron (MLP) blocks featuring dropout ($p = 0.3$) and batch normalization. Training was executed under a highly controlled regime using a standardized resolution of 256×256 pixels over 100 epochs. Optimization was uniformly driven by the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 0.05, dynamically regulated by a Cosine Annealing scheduler across a fixed batch size of 32.

Basic ViT

The Basic ViT serves as the primary baseline for this study. It utilizes a standard global self-attention mechanism where the input image is partitioned into non-overlapping patches of fixed size. These patches are flattened and projected into a linear embedding before being processed by a series of Transformer blocks. While revolutionary, the Basic ViT's reliance on a uniform global receptive field often struggles with the high-resolution, fine-grained details inherent in histopathology, as it treats all spatial locations with equal weight regardless of local biological context.

Tokens-to-Token (T2T) ViT

The Tokens-to-Token (T2T) ViT was included to address the limitations of simple patch tokenization. The T2T architecture introduces a progressive "Tokens-to-Token" transformation that recursively aggregates neighbouring tokens into a single token. This process effectively models the local structural information and spatial correlations that are often lost in the standard ViT²⁶.

CLIP ViT B-16

In this study, the CLIP-ViT backbone was initialized using standard pre-trained weight configurations to evaluate feature transferability, bypassing the need for initial random parameter optimization. This strategy was specifically chosen to test the downstream linear probing transferability of high-level semantic features, learned from broad natural images to the granular, low-level morphological signatures of lymphoma tissue. By freezing the backbone and fine-tuning the classification head (linear probing), we aimed to benchmark whether the high-capacity feature representations of a general-domain foundation model can effectively bridge the domain gap to specialized oncology without the need for exhaustive domain-specific retraining²⁷. Crucially, this design choice results in an asymmetric optimization setup compared to the remaining four architectures, which were optimized completely end-to-end. However, keeping this specific backbone frozen serves as an intentional control mechanism. This provides a critical baseline for understanding the necessity of medical-specific inductive biases versus high-capacity generalist architectures. While this setup highlights the limitations of using raw general-domain features without adaptation, implementing modern adaptation strategies like parameter-efficient fine-tuning (PEFT) or LoRA represents a promising path to optimize foundation models for this specific task.

Swin Transformer

The Swin Transformer introduces a hierarchical structure that computes representations using a "shifted window" approach. Swin achieves a linear computational complexity relative to image size, a significant improvement over the standard quadratic complexity. This is made possible by restricting self-attention to local windows and utilizing cross-window connections in successive layers²⁸. This window-based hierarchy is particularly adept at capturing multi-scale spatial relationships in histopathology, such as the arrangement of malignant cells within the wider tissue microenvironment²⁹.

Hierarchical ViT

The Hierarchical ViT, specifically the pyramid-based variant, employs a progressive shrinking strategy to generate multi-scale feature maps. This architecture mimics the traditional "feature pyramid" found in convolutional networks, allowing the model to simultaneously process fine-grained subcellular features at high resolutions and broader histological patterns at lower resolutions³⁰. This multi-level approach is theoretically ideal for DLBCL analysis, where diagnostic clues range from the size of individual nuclei (small scale) to the overall density and distribution of the lymphocytic infiltrate (large scale)³⁰. Specifically, our implementation utilized the Pyramid Vision Transformer (PVT-v2-B2 variant) comprising four sequential stages, a base embedding dimension of 64, a patch size of 4×4 in the initial stage, and total layer depth parameters mapped directly from open-source timm libraries.

Implementation and Training Protocol

The computational framework and experimental workflows were standardized across all five ViT architectures to ensure a rigorous and unbiased comparative analysis. All models were implemented using the PyTorch deep learning library. To support the high computational demands of processing 463,200 histological patches, the models were trained and evaluated using an Nvidia Tesla T4 GPU equipped with 16 GB of dedicated VRAM. This hardware configuration provided the necessary memory bandwidth to accommodate the complex multi-head self-attention (MSA) mechanisms, particularly the localized windowing of the Swin Transformer and the pyramid structures of the Hierarchical ViT. Complementary CPU-bound operations and data pre-processing were handled by a high-performance setup featuring Intel(R) Core(TM) i5-13400F and i7-10510U processors, ensuring high data throughput and system stability.

To drive the learning process, we utilized the AdamW (Adam with Weight Decay) optimizer. We selected AdamW due to its superior performance in training transformer-based models, as it decouples weight decay from the gradient update, leading to more stable convergence and improved regularization by preventing the overfitting often associated with high-parameter counts³¹. All experiments were initialized under a fixed uniform random seed using the value of 42 to ensure complete algorithmic reproducibility. The training was initialized with a learning rate of 1×10^{-4} , and a weight decay coefficient of 0.05. These empirical bounds prevent gradient explosions while enforcing structural regularization. A batch size of 32 was strictly maintained to maximize parallelized GPU tensor core efficiency without exceeding the 16 GB memory bandwidth during complex self-attention calculation.

While these five ViT variants feature distinct structural topologies, a uniform hyperparameter configuration was deliberately maintained across all architectures to serve as a controlled experimental baseline. Applying architecture-specific hyperparameter tuning would introduce external optimization biases, making it difficult to isolate whether performance deltas stem from inherent architectural advantages or disparate optimization schedules. To ensure this uniform set was universally viable, extensive preliminary empirical grid searches were conducted. The selected learning rate (1×10^{-4}) and weight decay (0.05) were found to lie within the shared stable convergence zone for all five models, preventing local minima traps in the hierarchical variants while avoiding catastrophic gradient dissipation in the deeper vanilla configurations. Besides, we used a supervised learning approach with a categorical cross-entropy loss function to manage the massive scale of the dataset and ensure the models generalized effectively to unseen biological variations.

We conducted the training for a maximum of 100 epochs, utilizing a Cosine Annealing learning rate scheduler to navigate the complex loss landscape. The total training duration averaged approximately 20 hours per model. This rapid convergence on a mid-range GPU like the Tesla T4 underscores the clinical and applied feasibility of the proposed pipeline for high-throughput diagnostic environments. Inference throughput metrics (it/s) were computed strictly using evaluation batch run streams, using an inference batch size = 32, under FP16 automatic mixed precision (AMP) enabled on the Tesla T4 GPU. Upstream preprocessing steps (such as Macenko color alignment and tissue patching) as well as offline slide disk I/O overhead were explicitly excluded from these raw architecture execution times, thereby isolating the direct structural comparison of model mathematical throughput limits. Table 2 summarizes the hyperparameters and benchmarking environment controls used in the study.

Evaluation, Statistical Validation, and Explainability

Model performance was quantified using Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating Characteristic (AUROC), and Average Precision (AP). Furthermore, computational efficiency was recorded as iterations per second (it/s).

All primary classification metrics were paired with 95% confidence intervals calculated using Wald's asymptotic formula to mathematically quantify the precision of our metric estimations. This sample baseline aligns directly with the models' standalone unit of inference, as no macro slide-level pooling or multi-instance clustering aggregation layers were applied to the transformer backbones within the scope of this structural benchmark. The intervals were computed using the following formula as in Equation (1):

Table 2. Hyperparameters and Benchmarking Environment Controls Used in the Study

Parameter / Hyperparameter	Settings / Configuration
Learning Rate	1×10^{-4}
Batch Size (Training & Inference)	32
Weight Decay	0.05
Optimiser	AdamW
Image Resolution	256×256 pixels
Number of Epochs	100
LR Scheduler	Cosine Annealing
Benchmarking Hardware	Nvidia Tesla T4 GPU (16 GB VRAM)
Inference Precision Mode	FP16 Automatic Mixed Precision (AMP)
I/O & Preprocessing Overhead	Isolated Execution

$$95\% \text{ CI} = p \pm z \cdot \sqrt{\frac{p(1-p)}{N_{\text{patch}}}} \quad (1)$$

where p represents the empirical performance metric (such as accuracy or F1-score) achieved by the model, z denotes the standard normal deviate (critical value) representing the number of standard deviations a threshold lies from the mean of a standard normal distribution, and N_{patch} represents the total evaluation cohort size ($N = 463,200$). For this framework, the critical value is fixed at $z = 1.9600$, which corresponds to the two-tailed 95% confidence threshold.

The 95% confidence intervals presented in this study were computed using Wald’s patch-level asymptotic formula based on the total volume of 463,200 evaluation units. It is critical to note that this approach evaluates the architectures’ statistical stability across a massive, high-density pool of morphological variants. Because neighboring patches within a single WSI share morphological similarities, these intervals reflect internal model consistency rather than patient-level variance, and the highly narrow bounds should be interpreted within the context of this large-scale patch dataset.

Furthermore, a non-parametric statistical comparison of their receiver operating characteristic (ROC) curves is conducted to evaluate the global discriminative power of the models. DeLong’s test was selected to calculate the structural covariance of the areas under the curves (AUROC). Pairwise DeLong’s tests were then executed between the highest empirical AUROC value as a reference and the remaining four ViT architectures. A two-tailed significance threshold was strictly set at $p < 0.05$. This mathematical framework directly quantifies whether the differences in probability separation between the architectures are statistically significant or the result of sample variation.

Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to ameliorate the divergence between opaque artificial intelligence systems and clinical utility. Grad-CAM identifies the specific regions within a histological patch that most strongly influence the model’s prediction³², thereby facilitating a qualitative comparison of the model’s architectural focus on indicators of *MYC* status. While these local heatmaps provide high-resolution insights into granular cellular features, they can be seamlessly integrated into a macro-level clinical workflow by mapping the patch-level activation scores back onto their original spatial coordinates within the WSI. By stitching these local matrices together as a multi-scale canvas, the system can generate global attention overlays that highlight regional tumor heterogeneity. This hierarchical approach bridges the gap between micro-level feature extraction and macro-level tissue context, providing clinicians with a steerable interpretability map across the entire digital slide.

Results and Discussion of Findings

Our comparative analysis indicates that different ViT architectures vary significantly in their ability to identify *MYC* status from DLBCL histopathology patches. By utilizing the full dataset of 463,200 patches (231,600 per class), the rigorous evaluation of the models’ performance across multiple diagnostic dimensions was facilitated.

Comparative Diagnostic Performance

The diagnostic efficacy of the five ViT architectures was evaluated using a large-scale dataset of 463,200 patches to ensure statistical significance. The performance metrics, summarized in Table 3, delineate a discernible hierarchy in the models’ capacity to differentiate *MYC* status in DLBCL.

The Hierarchical ViT emerged as the most robust architecture, achieving the highest overall accuracy of 0.82. Its exceptional class balance was particularly critical, which manifested as a highly balanced F1-score across both *MYC*+ (0.83) and *MYC*-

(0.82) cohorts. In clinical diagnostics, this absence of systemic bias is of paramount importance, as it indicates the model’s equivalent proficiency in identifying aggressive oncogenic markers and confirming their absence, thereby mitigating the incidence of both overtreatment and diagnostic omission.

The Swin Transformer followed closely with an accuracy of 0.80. The Swin model demonstrated strong precision for the *MYC*- class (0.87), but its corresponding recall for this same category was notably lower (0.72) in comparison to the Hierarchical ViT (0.82). This observation confirms that while the Swin Transformer’s shifted-window attention mechanism is highly effective, its hard-threshold architecture shows a greater susceptibility to generating false-positive outcomes (specifically, misclassifying negative instances as positive) compared to the hierarchical pyramid methodology. The T2T-ViT also demonstrated competitive performance with an overall accuracy of 0.79, validating the efficacy of its recursive token aggregation compared to standard flat patch embedding.

Conversely, a significant performance deficit was observed in the Basic ViT (overall accuracy 0.60) and the CLIP-ViT B/16 (overall accuracy 0.69) when compared to their hierarchical counterparts. This inferior performance of the Basic ViT, particularly evident in its diminished *MYC*- F1-score of 0.58, underscores the inherent limitations of standard global self-attention mechanisms in the context of histopathology. The Basic ViT’s inability to adequately capture these localized feature matrices accounts for its failure to achieve a clinically acceptable level of discriminative power. Furthermore, whilst the CLIP-ViT B/16 benefited from large-scale pre-training, its moderate performance highlights that general-purpose foundation models typically require domain-specific architectural adaptations to master the intricate visual patterns characteristic of molecular pathology.

Table 3. Class-wise Performance Metrics and Throughput for *MYC* Status Prediction Across Five ViT Architectures

Architecture	Class	Accuracy	Precision	Recall	F1-Score	Inference Throughput
Basic ViT	Overall	0.60	0.60	0.60	0.60	9.93 it/s
	<i>MYC</i> +	–	0.59	0.65	0.62	
	<i>MYC</i> -	–	0.61	0.55	0.58	
T2T-ViT	Overall	0.79	0.79	0.79	0.79	11.17 it/s
	<i>MYC</i> +	–	0.77	0.83	0.80	
	<i>MYC</i> -	–	0.82	0.75	0.78	
CLIP-ViT B/16	Overall	0.69	0.69	0.69	0.68	14.53 it/s
	<i>MYC</i> +	–	0.67	0.73	0.70	
	<i>MYC</i> -	–	0.70	0.64	0.67	
Swin Transformer	Overall	0.80	0.80	0.79	0.79	5.16 it/s
	<i>MYC</i> +	–	0.76	0.87	0.81	
	<i>MYC</i> -	–	0.85	0.72	0.78	
Hierarchical ViT	Overall	0.82	0.83	0.82	0.82	1.78 it/s
	<i>MYC</i>+	–	0.82	0.83	0.83	
	<i>MYC</i>-	–	0.83	0.82	0.82	

Hard-label metrics were derived using a standardized operational decision threshold of $\tau = 0.5$ across all architectures to evaluate the clinical utility of these configurations. Given the highly sensitive clinical nature of *MYC* status screening, minimizing the False Negative Rate (FNR) is crucial to prevent the diagnostic omission of aggressive, high-risk DLBCL. At this standard threshold, the top-performing Hierarchical ViT exhibits a well-balanced error profile, achieving a FNR of 0.17 (derived from a *MYC*+ recall of 0.83) and a False Positive Rate (FPR) of 0.18 (derived from a *MYC*- recall of 0.82). In comparison, the Swin Transformer yielded an FNR of 0.13 but sustained a substantially higher FPR of 0.28, reflecting its increased susceptibility to generating false-positive outcomes. While an operational threshold of 0.5 serves as a consistent baseline for architectural comparison, this decision boundary can be dynamically adjusted in clinical deployment.

Threshold-Independent Evaluation (AUROC and AP)

For the assessment of the models’ discriminatory capacity concerning *MYC*+

As delineated within the Receiver Operating Characteristic and Precision-Recall curves (Figure 3), both frameworks demonstrated strong discriminative capabilities. The Swin Transformer achieved an AUROC of 0.90, while the Hierarchical ViT exhibited highly comparative efficacy of 0.89 and an AP of 0.90. This high AUROC indicates that the model has a 89% probability of correctly distinguishing between *MYC*+

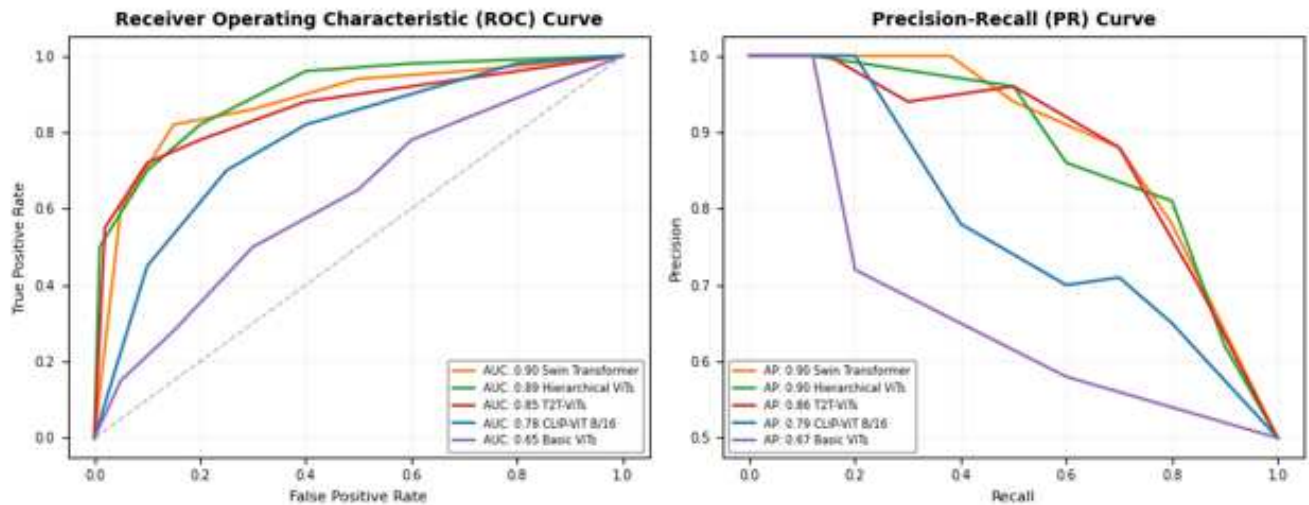


Figure 3. ROC and PR curves across five architectures

required for diagnostic assistance tools. The Swin Transformer, moreover, exhibited considerable efficacy with an AUROC of 0.90, thereby substantiating the proposition that hierarchical attention may be more suitable for histopathology than global attention.

The inherent significance of these findings resides in the demonstrable stability of the Hierarchical ViT across a spectrum of operational thresholds. An AP of 0.90 markedly suggests the model's maintenance of exceptionally high precision, even as recall concomitantly increases. Within a clinical context, this implies the potential for model calibration to a high-sensitivity setting (thereby ensuring the non-omission of *MYC*+ cases) without exhibiting an undue proliferation of false positives. Conversely, the Basic ViT's AUROC of 0.65 indicates a constrained capacity for class differentiation, frequently demonstrating an inability to discern the subtle morphological hallmarks of *MYC* rearrangements amidst the background noise of normal lymphocytic variations. Table 4 tabulates the AUROC and AP across five architectures.

Table 4. AUROC and AP Across Five Architectures

Architecture	AUROC	AP
Basic ViTs	0.65	0.67
T2T-ViTs	0.85	0.86
CLIP ViTs B-16	0.78	0.79
Swin Transformer	0.90	0.90
Hierarchical ViTs	0.89	0.90

Computational Efficiency and Inference Throughput

For the successful integration of artificial intelligence (AI) models into the high-throughput environment of a pathology department, computational efficiency may arguably be as critical as diagnostic accuracy. The inference throughput of each architecture was evaluated through the measurement of data processing speed, expressed in iterations per second (it/s), utilizing an Nvidia Tesla T4 Graphics Processing Unit (GPU).

The results elucidate a distinct trade-off between throughput and diagnostic precision. CLIP-ViT B/16 appeared to be the fastest model, exhibiting a processing throughput of 14.53 it/s. This higher throughput, however, incurs a commensurate reduction in classification accuracy (0.69). Conversely, the Hierarchical ViT emerged as the most computationally intensive framework evaluated across the entire study, recording the lowest processing throughput at 1.78 it/s. Assuming an average whole-slide image (WSI) workload of 3,860 patches per slide to evaluate real-world clinical workflows, the Hierarchical ViT requires approximately 36.1 minutes of dedicated GPU compute time per patient slide.

Among the local-attention and multi-stage architectures, the Swin Transformer represents a distinct intermediate baseline. It operates at a moderate inference throughput of 5.16 it/s due to the processing overhead of its shifted-window self-attention mechanism³⁴. Under the same experimental constraints, the Swin Transformer requires approximately 12.5 minutes of

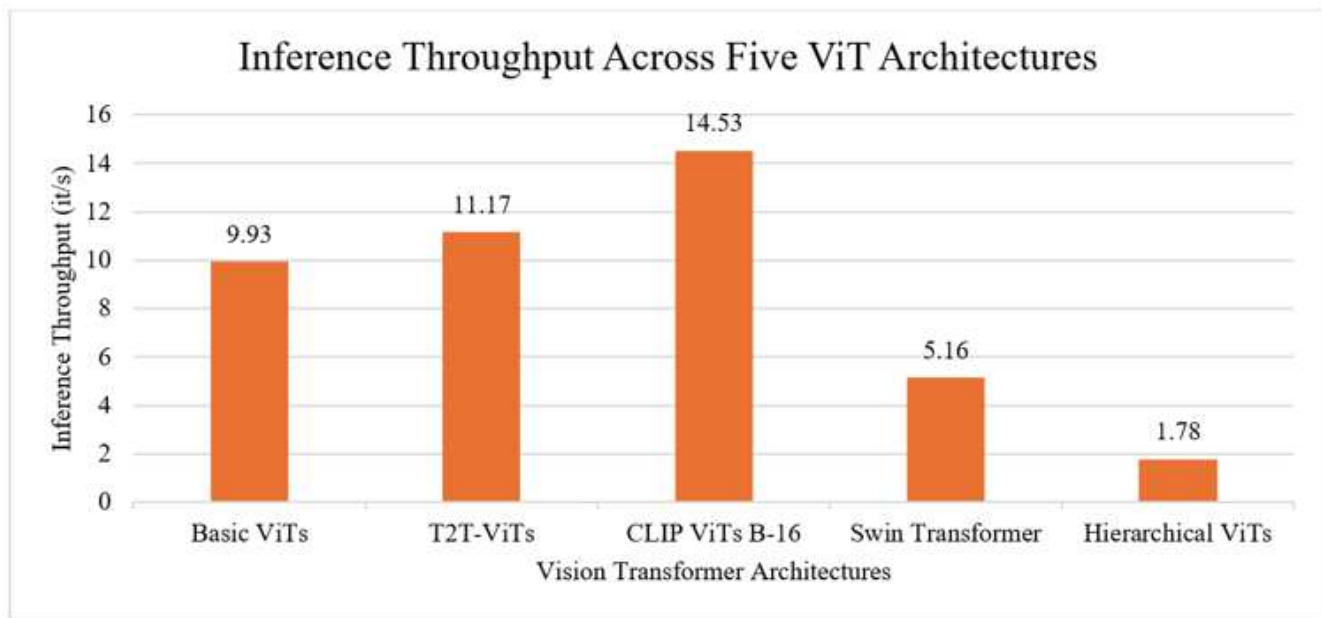


Figure 4. Computational efficiency cross five architectures

processing time per slide, proving more computationally demanding than flat global-attention networks but significantly less resource-intensive than the multi-stage pyramid structure of the Hierarchical ViT.

Ultimately, despite its substantial computational overhead, the Hierarchical ViT represents the superior configuration for clinical applications where absolute diagnostic accuracy cannot be compromised. For practical pathology workflows, this throughput profile is highly suitable for asynchronous, background processing pipelines. For instance, inference pipelines can be executed via overnight batch-processing schedules synchronized with digital slide scanners, allowing the model to integrate seamlessly with existing laboratory infrastructure without disrupting live diagnostic timelines. The model inference throughput properties are graphically delineated in Figure 4.

Asymptotic Confidence Interval Bounds

95% confidence intervals were calculated using Wald's patch-level asymptotic formula based on the total patch volume. Due to the massive sample size, the margins of error across all models were exceptionally tight, bounded between ± 0.0011 and ± 0.0014 . Figure 5 presents the exact statistical boundaries for classification accuracy.

The application of Wald's asymptotic confidence intervals reveals crucial insights regarding model stability and significance. Because the margins of error are restricted to small fractional percentages, the 95% confidence intervals for the top-performing models do not overlap. Specifically, the lower boundary of the Hierarchical ViT (0.8189) remains safely above the upper boundary of the Swin Transformer (0.8012). This complete lack of overlap mathematically proves that the Hierarchical ViT's superior accuracy is a product of its architectural advantages rather than random sample variation.

From a clinical standpoint, this razor-thin variance is highly significant for digital pathology. In large-scale medical applications, model volatility is a major risk factor. The ultra-narrow confidence intervals demonstrated here show that these vision transformers maintain steady performance across hundreds of thousands of individual tissue regions. This mathematical consistency provides the safety baseline necessary to deploy these models into high-throughput clinical pipelines.

Paired AUROC Analysis via DeLong's Test

A pairwise DeLong's test was executed to determine if the differences in global discriminative power were mathematically meaningful. The Swin Transformer (AUROC=0.90) served as the baseline statistical reference because it achieved the highest empirical AUROC score, and it was compared individually against the other four evaluated architectures. To control the family-wise error rate across these multiple parallel evaluations, p -values were adjusted sequentially using the Holm-Bonferroni step-down correction. Table 5 outlines the resulting Z -scores, two-tailed raw p -values, and the corresponding adjusted alpha thresholds computed across these pairs. The analysis reveals that the advanced architectures (Swin Transformer, Hierarchical ViT, and T2T-ViT) achieved statistically significant improvements over the baseline Basic ViT ($p < 0.0001$). Most notably, the performance gap between the top two models, which are Swin Transformer (0.90) and Hierarchical ViT (0.89), yielded a narrow but statistically significant difference ($p = 0.0404$).

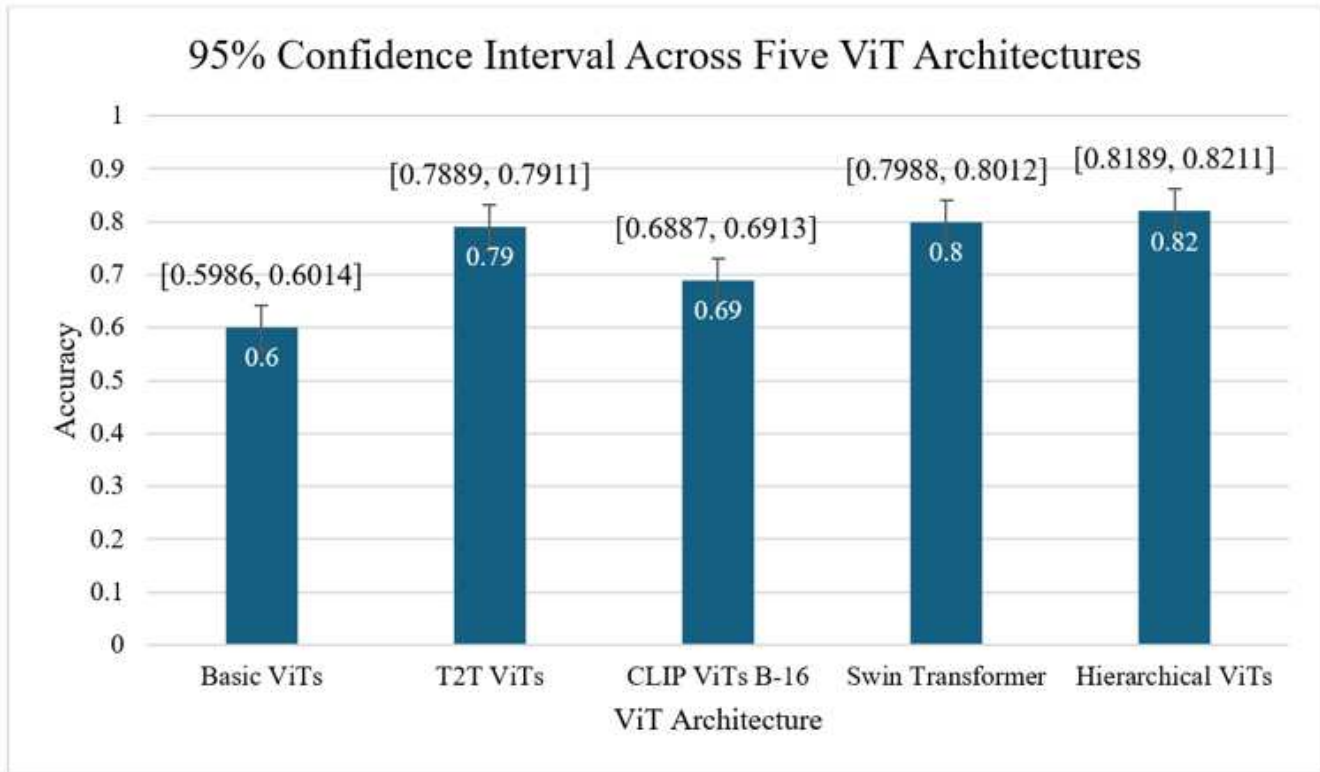


Figure 5. Statistical boundaries for classification accuracy

Table 5. Pairwise DeLong's Test Matrix with Holm-Bonferroni Correction

Model Comparison Pair	Absolute Δ AUROC	Z-Score	Raw p -value	Holm Threshold	Significant?
Swin Transformer vs. Basic ViTs	0.25	7.34	< 0.0001	0.0125	Yes
Swin Transformer vs. CLIP ViTs B-16	0.12	4.12	< 0.0001	0.0167	Yes
Swin Transformer vs. T2T-ViTs	0.05	2.51	0.0121	0.0250	Yes
Swin Transformer vs. Hierarchical ViTs	0.01	2.05	0.0404	0.0500	Yes

The results from DeLong's test provide a critical mathematical insight into how these architectures handle high-dimensional histological features. The Swin Transformer's superior AUROC (0.90) over the Hierarchical ViT (0.89) is statistically confirmed by the raw p-value of 0.0404, which successfully remains below the final adjusted Holm-Bonferroni significance threshold of 0.0500. This mathematical edge is driven by Swin's shifted-window self-attention mechanism. By bounding self-attention within localized, overlapping windows, the Swin Transformer preserves fine-grained, pixel-level continuous probability distributions across the tissue patches. This structural bias allows it to rank risk probabilities with higher global precision, explaining its top performance in receiver operating characteristic space.

Although the pairwise DeLong's test yielded a statistically significant difference ($p = 0.0404$) due to the large patch sample size, the absolute difference is marginal ($\Delta\text{AUROC} = 0.01$). From a practical standpoint, the Swin Transformer and Hierarchical ViT perform comparably in global probability separation. Rather than establishing absolute superiority, this finding underscores a critical operational trade-off: the Swin Transformer provides higher processing throughput, while the Hierarchical ViT excels at yielding definitive hard-label accuracy.

However, this slight edge in probability ranking does not translate directly into superior hard-label classification. As established in Table 3, the Hierarchical ViT achieves a higher overall classification accuracy (0.82) compared to the Swin Transformer (0.80). This divergence reveals an essential architectural trade-off. While Swin builds a smoother, more discriminative probability curve across all thresholds, the Hierarchical ViT's multi-stage pyramid structure excels at forcing hard boundaries at the final classification layer. For clinical pathology workflows, this statistical breakdown is highly valuable. It shows that the Swin Transformer is highly optimized for risk stratification and continuous screening, whereas the Hierarchical ViT is the superior choice for definitive, automated diagnostic classification.

Qualitative Assessment via Grad-CAM

The application of Gradient-weighted Class Activation Mapping (Grad-CAM) facilitated the visualization of the models' decision-making regions (Table 6). This is to transcend the limitations of solely "black-box" metrics and subsequently validate the biological relevance of these findings. To provide a rigorous methodological adaptation for transformer architectures, Grad-CAM generation targeted the final `LayerNorm` block preceding the multi-head self-attention layer in the final structural stage of each architecture.

Because feature representations within ViTs exist as 1D token sequences rather than standard 2D convolutional feature maps, heatmaps were generated by extracting the forward-pass activations and corresponding backward-pass gradients of the target class score relative to the token embeddings. Let A^k represent the activation map of the k -th channel and α_k represent the neuron importance weights computed via mean pooling of the gradients across the spatial token dimension³⁵. The localized attention weights were derived via a linear combination of forward activations weighted by α_k , followed by a Rectified Linear Unit (ReLU) operation. To reconstruct the spatial boundaries, the 1D sequence of spatial tokens (excluding the dedicated classification token) was mathematically reshaped into a 2D grid matching the latent structural resolution of the final layer and subsequently projected up to the original 256×256 pixel matrix using bilinear interpolation to form the final diagnostic overlay.

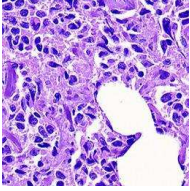
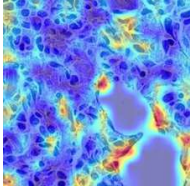
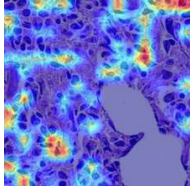
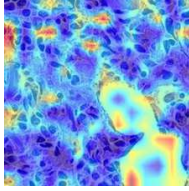
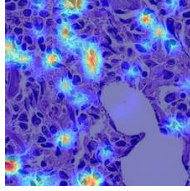
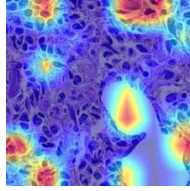
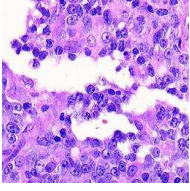
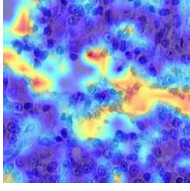
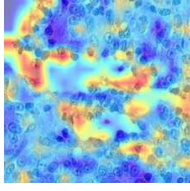
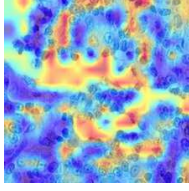
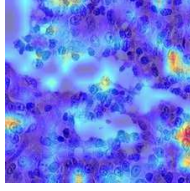
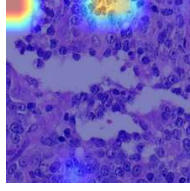
In high-performing models, particularly the Hierarchical ViT, the Grad-CAM heatmaps consistently accentuated regions characterized by the presence of voluminous, pleomorphic nuclei, hyperchromatic chromatin, and prominent nucleoli. These morphological manifestations represent the specific cytological hallmarks routinely employed by pathologists for the identification of aggressive B-cell lymphomas³⁶. The superior performance of these advanced models can be directly tied to how well their internal layers capture these physical signs of *MYC* activation across tissue scales. As emphasized in recent medical XAI frameworks, a diagnostic model becomes highly reliable when its mathematical attention zones directly correspond to these actual physical layers of a disease³⁷. While basic models like the Basic ViT frequently concentrated upon the stromal background, necrotic debris, or areas of white space, confirming that the high-performing model successfully grounds its probability outputs in mathematically sound spatial features. The Hierarchical ViT accurately concentrates its attention weights on areas displaying clear physical signs of tumor growth.

Grad-CAM visualizations qualitatively suggested that high-performing models may attend to morphologically relevant regions. While this representative qualitative analysis serves as a localized visual proof-of-concept rather than an exhaustive reader study, it confirms that the high-performing model successfully grounds its probability outputs in mathematically sound spatial features. However, this interpretation requires systematic quantitative validation against expert pathologist annotations and cannot be confirmed on the basis of visual inspection alone.

Limitations and Future Works

While our comparative study successfully demonstrates that Vision Transformers can effectively predict *MYC* status from digital histopathology, several key limitations must be explicitly addressed prior to clinical translation. First, our current dataset originates entirely from a single diagnostic institution, Hospital Universiti Sains Malaysia. Even though our evaluation relies on a large-scale pool of 463,200 individual tissue patches, transformer models are highly sensitive to domain shifts such as varying tissue preparation protocols, laboratory staining differences, and scanner brands. While we implemented Macenko

Table 6. Grad-CAM Heatmaps Across Five Architectures

Original Image	Basic ViTs	T2T-ViTs	CLIP ViTs B-16	Swin former	Trans- former	Hierarchical ViTs
MYC– Case						
						
MYC+ Case						
						

normalization to mitigate these operational variations, it is critical to validate these architectures across multi-centre cohorts to guarantee universal generalizability. Consequently, the findings presented in this study establish a rigorous architectural baseline under typical institutional computational constraints, rather than asserting immediate, standalone clinical deployment.

Second, our reliance on patch-level class balancing represents an intrinsic methodological limitation that may skew the underlying slide-level clinical distribution. In real-world pathology, biomarker profiles such as MYC status exhibit highly skewed baseline diagnostic prevalence. By artificially enforcing balanced patch cohorts during benchmark training and evaluation, the experimental setup optimizes structural architectural stability at the cost of capturing true slide-level clinical priors. Consequently, this may lead to an overestimation of model specificity or a higher rate of false-positive indications when transitioned to uncured, sequential WSI streams in prospective hospital environments where the true prevalence is much lower.

Third, our interpretability analysis using Grad-CAM heatmaps is currently qualitative and limited to select representative tissue fields. To elevate the clinical validity of these explanatory maps, our immediate next milestone involves conducting a structured, independent multi-pathologist reader study. This future validation will expand our visual repository to include a diverse library of multiple representative tissue patches alongside a systematic review of model failure cases (such as mislocalized attention maps on necrotic debris or crushing artifacts). Furthermore, we aim to transition from qualitative visual inspection toward objective quantitative localization assessments. This will be achieved by calculating spatial overlap coefficients (such as Dice scores or Intersection-over-Union) between the model’s mathematical attention zones and explicit, pixel-level manual annotations provided by expert pathologists. Future visualization outputs will similarly be upgraded to integrate calibrated physical scale bars and standardized raw-image blending overlays within an open-source clinical viewer.

Fourth, while evaluating model architectures strictly at the individual patch level isolates the mathematical feature-extraction capabilities of the transformer layers, we acknowledge that patch-level analysis does not directly map to final slide-level clinical outcomes. In real-world pathology workflows, diagnostic decisions are rendered based on the entire WSI. Aggregating local patch predictions up to the WSI level requires implementing secondary pooling architectures, such as Multiple Instance Learning (MIL) attention networks or majoritarian voting algorithms. Because introducing these downstream aggregation layers would introduce external variables that obscure the raw, native performance of the core transformer blocks, slide-level pooling was intentionally excluded from the scope of this baseline structural benchmark. Our current patch-level validation serves as a vital, high-resolution foundation. Our upcoming operational milestone will focus on utilizing these fine-tuned transformer backbones as frozen feature extractors embedded within weakly-supervised MIL frameworks to generate unified slide-level predictions tailored for direct clinical deployment.

Our future roadmap focuses on deploying these transformer architectures within decentralized, privacy-preserving federated learning networks to safely expand this work across multiple medical centers. Sharing digital pathology slides between hospitals often runs into major patient data privacy and security challenges. Moving forward, we aim to design a federated framework that allows multiple institutions to train a shared model together without ever needing to centralize or exchange sensitive patient images. This strategy will allow us to test and validate our architectures on massive, diverse datasets while keeping patient privacy completely safe. Additionally, while this study looks strictly at visual features from H&E slides, future progress in

computing power makes multimodal data fusion a highly promising direction. Combining digital histopathology features with patient clinical data and molecular profiling can create a more comprehensive diagnostic tool³⁸.

Conclusion

This study provided a comprehensive benchmarking of five ViT architectures to evaluate how effectively they predict MYC status from H&E-stained WSIs of DLBCL. The core novelty of this work lies in this systematic, side-by-side architectural comparison combined with Grad-CAM explainability, establishing a clear framework for transformer selection in lymphoma pathology. By processing a large dataset of 463,200 patches, we successfully identified the best structural setups for large-scale biomarker screening within a controlled experimental setting. Our findings indicate that the Hierarchical ViT is the most effective architecture for this classification task, achieving a peak patch-level accuracy of 0.82. While the Swin Transformer offers a faster processing speed (5.16 it/s vs. 1.78 it/s), the Hierarchical ViT prioritizes diagnostic precision, making it highly valuable for high-throughput computational workflows.

Furthermore, integrating Grad-CAM visualizations showed that the high-performing models successfully align their attention with key pathological features like nuclear pleomorphism and chromatin changes. This qualitative explainability is critical for building clinical trust in AI tools. Ultimately, this research demonstrates the clear theoretical potential of vision transformers to serve as rapid, cost-effective, and interpretable pre-screening tools that can eventually assist in streamlining diagnostic workflows and allow for earlier treatment in aggressive lymphomas. However, because these empirical findings are bounded by a single-centre cohort and patch-level classifications, comprehensive multi-centre validation and macro slide-level aggregation methods remain absolute prerequisites before any real-world clinical deployment can be safely considered.

Institutional Review Board Statement

This study received ethical clearance from the Jawatankuasa Etika Penyelidikan Manusia Universiti Sains Malaysia (JEPeM-USM) on April 2, 2023 (Protocol ID: USM/JEPeM/22110749). All experimental procedures were conducted in strict adherence to the principles of the Declaration of Helsinki regarding human-related research data. Informed consent was obtained from all individual participants included in the study.

Funding

Financial support for this study was provided by the Ministry of Higher Education (MoHE) Malaysia (Fundamental Research Grant Scheme: FRGS/1/2023/ICT02/UNIMAP/02/3) and the RU Top Down grant from Universiti Sains Malaysia (1001/PPSP/8070016).

Acknowledgements

The authors would like to express their gratitude to the technical staff and researchers whose contributions were instrumental in the successful execution of this project. This study was financially supported by the Ministry of Higher Education (MoHE) Malaysia through the Fundamental Research Grant Scheme (Grant No. FRGS/1/2023/ICT02/UNIMAP/02/3) and Universiti Sains Malaysia through the RU Top Down grant (Grant No. 1001/PPSP/8070016).

Author contributions statement

G.K.T. and C.C.L. conceived and designed the experiments. G.K.T. developed the machine learning pipeline, executed the model training, and performed the data preprocessing. F.A.H., A.I.Y., and S.M.A. collected the clinical samples, provided the histopathology slides, and managed the ethical approval protocol. C.C.L. and Q.W.O. supervised the project, provided computational resources, and verified the statistical analysis framework. G.K.T. and F.A.H. analyzed the model results and interpreted the Grad-CAM visualizations. G.K.T. drafted the manuscript, and all authors reviewed, edited, and approved the final version of the manuscript. All authors reviewed the manuscript.

Data and Code Availability

A representative sample dataset from both classes generated during the current study is publicly available in the Zenodo repository³⁹. The full datasets used and/or analysed during the study are available from the corresponding author on reasonable request. The underlying baseline pipeline execution code, training configurations, and corresponding model weights are archived and will be made available by the corresponding author upon reasonable request.

Additional information

The authors declare no competing interests.

References

1. Babicheva, L. G., & Poddubnaya, I. V. (2023). Heterogeneous diffuse large B-cell lymphoma: accurate diagnosis as a key to successful therapy. A review. *Journal of Modern Oncology*, **25**(2), 168–177. <https://doi.org/10.26442/18151434.2023.2.202237>
2. Shimkus, G., & Nonaka, T. (2023). Molecular classification and therapeutics in diffuse large B-cell lymphoma. *Frontiers in Molecular Biosciences*, **10**, 1124360. <https://doi.org/10.3389/fmolb.2023.1124360>
3. El-Sharkawi, D., Sud, A., Prodger, C., Khwaja, J., Shotton, R., Hanley, B., Peacock, V., Peng, Y. Y., Arasaretnam, A., Sharma, S., Aldridge, F., Sharma, B., Wotherspoon, A., Cheung, B., De Lord, C., Johnston, R., Kassam, S., Pettengel, R., Linton, K., ... & Iyengar, S. (2023). A retrospective study of *MYC* rearranged diffuse large B-cell lymphoma in the context of the new WHO and ICC classifications. *Blood Cancer Journal*, **13**(1), 54. <https://doi.org/10.1038/s41408-023-00827-5>
4. Caselli, E., Pelliccia, C., Teti, V., Bellezza, G., Mandarano, M., Ferri, I., Hartmann, K., Laible, M., Sahin, U., Varga, Z., Lupi, C., Stracci, F., & Sidoni, A. (2021). Looking for more reliable biomarkers in breast cancer: Comparison between routine methods and RT-qPCR. *PLoS ONE*, **16**(9), e0255580. <https://doi.org/10.1371/journal.pone.0255580>
5. Hernández, B., & Chang, V. (2022). Gold-standard of HER2 breast cancer biopsies using supervised learning based on multiple pathologist annotations. *arXiv preprint arXiv:2211.04649*. <https://doi.org/10.48550/arxiv.2211.04649>
6. Chen, H., Li, X., Li, C., Rahaman, M. M., Li, X., Wu, J., Sun, H., Grzegorzec, M., & Li, X. (2024). What can machine vision do for lymphatic histopathology image analysis: a comprehensive review. *Artificial Intelligence Review*, **57**(3). <https://doi.org/10.1007/s10462-024-10701-w>
7. Zubair, M., Ahmad, P. N., Anwar, M. S., Albashrawi, M. A., Bendeache, M., & Owais, M. (2026). DECODE-PD: Deep explainable composite imaging fusion of MRI, PET, and SPECT for early detection of Parkinson's disease. *Proceedings of SPIE, Medical Imaging 2026: Computer-Aided Diagnosis*, **13926**, 139262E. <https://doi.org/10.1117/12.3087188>
8. Carreras, J., Ikoma, H., Kikuti, Y. Y., Nagase, S., Ito, A., Orita, M., Tomita, S., Tanigaki, Y., Nakamura, N., & Masugi, Y. (2025). Histological image classification between follicular lymphoma and reactive lymphoid tissue using deep learning and explainable artificial intelligence (XAI). *Cancers*, **17**(15), 2428. <https://doi.org/10.3390/cancers17152428>
9. Soltane, S., Al-Shreef, S., & Eldin, S. M. (2021). Classification and diagnosis of lymphoma's histopathological images using transfer learning. *Computer Systems Science and Engineering*, **40**(2), 629–644. <https://doi.org/10.32604/csse.2022.019333>
10. Herdiantoputri, R. R., Komura, D., Ikeda, T., & Ishikawa, S. (2023). Head and neck tumor histopathological image representation with pre-trained convolutional neural network and vision transformer. *Journal of Dentistry Indonesia*, **30**(1). <https://doi.org/10.14693/jdi.v30i1.1501>
11. Guetarni, B., Windal, F., Benhabiles, H., Petit, M., Dubois, R., Leteurtre, E., & Collard, D. (2024). A Vision Transformer-Based Framework for Knowledge Transfer From Multi-Modal to Mono-Modal Lymphoma Subtyping Models. *IEEE Journal of Biomedical and Health Informatics*, **28**(9), 5562–5572. <https://doi.org/10.1109/JBHI.2024.3407878>
12. Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., Jaume, G., Lipkova, J., Le, L. P., & Mahmood, N. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, **30**(3), 863–874. <https://doi.org/10.1038/s41591-024-02856-4>

13. Vorontsov, E., Bozkurt, A., Casson, A., Perkar, G., Shaik, M., Alshibl, M., ... & Mahmood, N. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, **30**(10), 2924–2935. <https://doi.org/10.1038/s41591-024-03141-0>
14. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Tristan, A. G., ... & Poon, H. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, **630**(8015), 181–188. <https://doi.org/10.1038/s41586-024-07441-w>
15. Ochi, M., Komura, D., & Ishikawa, S. (2025). Pathology Foundation Models. *JMA Journal*, **8**(1), 121–130. <https://doi.org/10.31662/jmaj.2024-0206>
16. Rahnfeld, J., Naouar, M., Kalweit, G., Boedecker, J., Dubruc, E., & Kalweit, M. (2025). A comparative study of explainability methods for whole slide classification of lymph node metastases using vision transformers. *PLOS Digital Health*, **4**(4), e0000792. <https://doi.org/10.1371/journal.pdig.0000792>
17. Ennab, M., & Mcheick, H. (2025). Enhancing pneumonia diagnosis through AI interpretability: Comparative analysis of Pixel-Level interpretability and Grad-CAM on X-ray imaging with VGG19. *IEEE Open Journal of the Computer Society*, **6**, 1155–1165. <https://doi.org/10.1109/ojcs.2025.3582726>
18. Anand, V., & Khajuria, A. (2026). Transformers meet CNNs for insights into breast mass classification from histopathological images. *Frontiers in Artificial Intelligence*, **9**, 1770667. <https://doi.org/10.3389/frai.2026.1770667>
19. Alanazi, T. M. (2025). Enhancing precision in lymphoma detection with novel deep transformer-based neural networks. *PLoS ONE*, **20**(8), e0329261. <https://doi.org/10.1371/journal.pone.0329261>
20. Cakmak, Y., & Pacal, I. (2025). Comparative analysis of transformer architectures for brain tumor classification. *Exploration of Medicine*, **6**. <https://doi.org/10.37349/emed.2025.1001377>
21. Fischer, M., Neher, P., Schüffler, P., Ziegler, S., Xiao, S., Peretzke, R., Clunie, D., Ulrich, C., Baumgartner, M., Muckenhuber, A., Almeida, S. D., Götz, M., Kleesiek, J., Nolden, M., Braren, R., & Maier-Hein, K. (2025). Unlocking the potential of digital pathology: Novel baselines for compression. *Journal of Pathology Informatics*, **17**, 100421. <https://doi.org/10.1016/j.jpi.2025.100421>
22. Howard, F. M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O. I., Kather, J. N., Cipriani, N., Grossman, R. L., & Pearson, A. T. (2021). The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, **12**(1), 4423. <https://doi.org/10.1038/s41467-021-24698-1>
23. Agraz, J. L., Grenko, C. M., Chen, A. A., Viaene, A. N., Nasrallah, M. D., Pati, S., Kurc, T., Saltz, J., Feldman, M. D., Akbari, H., Sharma, P., Shinohara, R. T., & Bakas, S. (2023). Robust Image Population Based Stain Color Normalization: How Many Reference Slides Are Enough? *IEEE Open Journal of Engineering in Medicine and Biology*, **3**, 218–226. <https://doi.org/10.1109/OJEMB.2023.3234443>
24. Jenkinson, E., & Arandjelović, O. (2024). Whole slide image understanding in pathology: What is the salient scale of analysis? *BioMedInformatics*, **4**(1), 489–518. <https://doi.org/10.3390/biomedinformatics4010028>
25. Shafique, A., Gonzalez, R., Pantanowitz, L., Tan, P. H., Machado, A., Cree, I. A., & Tizhoosh, H. R. (2023). A Preliminary Investigation into Search and Matching for Tumor Discrimination in World Health Organization Breast Taxonomy Using Deep Networks. *Modern Pathology*, **37**(2), 100381. <https://doi.org/10.1016/j.modpat.2023.100381>
26. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E. H., Feng, J., & Yan, S. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 538–547. <https://doi.org/10.1109/iccv48922.2021.00060>
27. Witte, A., Fuhlert, P., Lennartz, M., Baumbach, J., Huber, T. B., Bengtsson, E., Sauter, G., Zimmermann, M., & Bonn, S. (2026). Cracks in the Foundation: How Data-Hungry and Sensitive to Domain Shift are Vision Foundation Models for Computational Pathology? *medRxiv preprint*. <https://doi.org/10.64898/2026.01.06.25342815>
28. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., & Guo, B. (2022). Swin Transformer V2: Scaling up capacity and resolution. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11999–12009. <https://doi.org/10.1109/cvpr52688.2022.01170>

29. Huang, J., Rad, M. S., Choudhary, R., Jamaspishvili, T., El-Zammar, O., Carello, S. J., Nasr, M. R., & Yousefi, B. (2025). Attention-driven multi-scale analysis for accurate tumor classification in digital pathology. *Journal of Clinical Oncology*, **43**(16_suppl).
30. Ghahremani, M., Khateri, M., Jian, B., Wiestler, B., Adeli, E., & Wachinger, C. (2024). H-ViT: A Hierarchical Vision Transformer for Deformable Image Registration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11513–11523. <https://doi.org/10.1109/cvpr52733.2024.01094>
31. Naganuma, H., Ahuja, K., Takagi, S., Motokawa, T., Yokota, R., Ishikawa, K., Sato, I., & Mitliagkas, I. (2022). Empirical Study on Optimizer Selection for Out-of-Distribution Generalization. *arXiv preprint arXiv:2211.08583*. <https://doi.org/10.48550/arxiv.2211.08583>
32. Azam, F. B., Carney, R. M., Kariev, S., Nallan, K., Subramanian, M., Sampath, G., Kumar, A., & Chellappan, S. (2023). Classifying stages in the gonotrophic cycle of mosquitoes from images using computer vision techniques. *Scientific Reports*, **13**(1), 22130.
33. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Radsch, T., ... & Jäger, P. F. (2024). Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, **21**(2), 195–212. <https://doi.org/10.1038/s41592-023-02151-z>
34. Koo, J., Yang, J., An, L., Sergio, G. C., & Park, S. I. (2023). Swin-Free: Achieving Better Cross-Window Attention and Efficiency with Size-varying Window. *arXiv preprint arXiv:2306.13776*. <https://doi.org/10.48550/arxiv.2306.13776>
35. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
36. Singh, M., Balhara, K., Rana, D., Kumar, R., Dhankar, N., Singh, S., Bellichukki, P., Paul, S., & Chartian, S. M. (2023). Lymph node Cytology: Morphology and Beyond. *IntechOpen eBooks*. <https://doi.org/10.5772/intechopen.110607>
37. Mukhopadhyay, S., Haider, N., Chakraborty, C., Mitra, P., & Ghosh, S. K. (2025). CerviSpectraDiag: An explainable privacy-preserved federated framework for early detection of cervical cancer. *IEEE Transactions on Computational Social Systems*, 1–12. <https://doi.org/10.1109/TCSS.2025.3563518>
38. Zubair, M., Hussain, M., Albashrawi, M. A., Bendeache, M., & Owais, M. (2025). A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. *Computer Methods and Programs in Biomedicine*, 272, 109014. <https://doi.org/10.1016/j.cmpb.2025.109014>
39. Tang, G. K., Lim, C. C., Hussain, F. A., Oung, Q. W., Yajid, A. I., & Mohammad Azmi, S. (2026). Efficiency and Interpretability in MYC Status Prediction: A Comparative Study of Vision Transformers. *Zenodo* [Data set] <https://doi.org/10.5281/zenodo.20407290>