Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Novel Hybrid Pyramid ViT-ConvFormer Architecture with Dynamic Early Stopping Analysis for Enhanced Diagnostic Efficiency in Diffuse Large B-Cell Lymphoma

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: Diffuse Large B-Cell Lymphoma, Digital Pathology, Early Stopping, MY C Status, PyConv

Posted Date: September 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10698568/v1

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

Additional Declarations: No competing interests reported.

A Novel Hybrid Pyramid ViT-ConvFormer Architecture with Dynamic Early Stopping Analysis for Enhanced Diagnostic Efficiency 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†], Sumayyah Mohammad Azmi[3,4†]

[1*]Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, Kampus Pauh Putra, Arau, 02600, Perlis, Malaysia.
[2*]Sport Engineering Research Centre, Universiti Malaysia Perlis, Kampus Pauh Putra, Arau, 02600, Perlis, Malaysia.
[3]Department of Pathology, School of Medical Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, 16150, Kelantan, Malaysia.
[4]Hospital Pakar USM, Kubang Kerian, 16150, Kelantan, Malaysia.
[5]Advanced Communication Engineering (ACE) Centre of Excellence, Universiti Malaysia Perlis, Arau, 02600, Perlis, Malaysia.

*Corresponding author(s). E-mail(s): geiki@studentmail.unimap.edu.my; cclim@unimap.edu.my;
Contributing authors: faezahtul@usm.my; qiwei@unimap.edu.my; aidy@usm.my; sumayyahazmi@moh.gov.my;
†These authors contributed equally to this work.

Abstract

Purpose: Accurate stratification of Diffuse Large B-Cell Lymphoma (DLBCL) into ***MYC*** subtypes is crucial for determining therapeutic intensity. While Vision Transformers (ViTs) excel at capturing global tissue context, their quadratic computational complexity results in slow convergence. To address this, we developed PyConv, a dual-stream network designed to accelerate training while maintaining precise DLBCL classification.

Methods: PyConv pairs local ConvFormer blocks with global Pyramid Vision Transformer (PVT) blocks, unifying cytological boundaries and macro-level tissue context via a Multi-Head Cross-Attention fusion layer. The architecture was evaluated using 180 whole-slide images (69,480 testing patches) against 13 state-of-the-art backbones.
Results: PyConv achieved a multi-class accuracy and macro F1-score of 0.9962, an absolute decision boundary for benign cohorts, and a malignant stratification AUROC of 0.9999. It reached optimal generalization by epoch 20, an 83.5% reduction in convergence time over pure transformers. While pure hybrid models showed slightly higher raw throughput, DeLong's testing confirmed performance variances were statistically non-significant. Grad-CAM, Integrated Gradient, and Attention Rollout verified that PyConv correctly prioritizes key cytological features like nuclear contours and prominent nucleoli.
Conclusion: PyConv offers a highly accelerated, computationally balanced alternative for DLBCL stratification without sacrificing diagnostic accuracy or clinical interpretability.

Keywords: Diffuse Large B-Cell Lymphoma, Digital Pathology, Early Stopping, ***MYC*** Status, PyConv

1 Introduction

Diffuse Large B-Cell Lymphoma (DLBCL) is an aggressive haematological malignancy characterized by pronounced clinical and biological heterogeneity. Accurate molecular stratification remains paramount in large B-cell lymphomas, where identifying MYC proto-oncogene rearrangements serves as a critical frontline prognostic indicator [1]. Patients presenting with MYC-positive ($MYC+$) DLBCL exhibit rapid disease progression and poor survival outcomes under standard R-CHOP chemoimmunotherapy [2], necessitating early detection to deploy intensive frontline regimens like dose-adjusted EPOCH-R [3]. However, the gold-standard diagnostic validation via Fluorescence in Situ Examination (FISH) break-apart assays are constrained by high institutional costs, labour-intensive protocols, and prolonged turnaround latencies [4], [5]. Automated patch networks help mitigate the issues of high inter-observer variability and slow optical review workflows during multi-scale Whole Slide Image (WSI) evaluations [6], [7], making digitized Haematoxylin and Eosin (H&E) screening an attractive alternative.

While Convolutional Neural Networks (CNNs) capture localized cytological features well, they struggle to model the broad spatial tissue relationships essential for comprehensive tissue profiling [8]. Conversely, pure Vision Transformers (ViTs) leverage Multi-Head Self-Attention (MHSA) mechanisms to capture global structures, but they suffer from high computational footprints [9], [10]. While pure self-attention layers excel at mapping broad context boundaries, they exhibit high training latencies and a characteristic "lazy learning" trajectory [11], [12]. This "transformer lag" causes slow optimization convergence and leaves models vulnerable to non-biological noise, such as staining variations [13], [14].

This study introduces PyConv, a novel hybrid deep learning architecture that splits representation learning into an asymmetrical, dual-stream topology driven by a multi-stage convolutional stem. The local token-mixing stream leverages cascaded ConvFormer blocks optimized with large-kernel depthwise-separable convolutions to capture fine-grained cytological boundaries with linear complexity. Concurrently, the global attention stream employs Pyramid Vision Transformer (PVT) blocks with asymmetric Spatial Reduction Attention (SRA) to map macro-level tissue distribution patterns across the stroma. The structural representations from both tracks are dynamically reconciled via a Multi-Head Cross-Attention (MHCA) fusion layer that imposes local anchors onto the global contextual landscape.

We evaluate our approach using a massive cohort of 694,800 histopathological patches generated from 180 high-resolution WSIs ($MYC+$, $MYC-$, and benign Reactive Lymph Node (RLN) controls). By systematically analysing early stopping trajectories across thirteen state-of-the-art backbones, we assess the trade-offs between parameter efficiency, convergence speed, and diagnostic accuracy under a strict Cosine Annealing optimization protocol. Finally, we implement multiple visual explainability paradigms, including Grad-CAM, Integrated Gradients, and Attention Rollout, to confirm that the network's internal representations tightly align with established pathological criteria, ensuring clinical transparency and interpretability for high-throughput oncology pipelines.

2 Related Works

2.1 The Evolution of Digital Pathology in Lymphoma Stratification

The clinical management of DLBCL has shifted toward precision molecular profiling, where the identification of MYC rearrangements is a critical prognostic indicator [1]. Traditionally, this stratification relied on manual immunohistochemistry (IHC) scoring, a process that is prone to significant inter-observer variability [15]. Recent advancements in whole slide imaging have enabled the application of deep learning to automate these workflows. Early diagnostic models primarily utilized CNNs owing to their inherent inductive bias and translation invariance, which are effective for capturing local cellular features, such as nuclear pleomorphism [16]. However, pure CNN architectures often struggle with the "global context" problem, that is, the inability to model long-range spatial relationships between distant cell clusters across a massive 256×256 patch [17].

2.2 Vision Transformers and Optimization Limitations

The ViT marked a paradigm shift in medical imaging by treating image patches as sequences, allowing global dependency modelling through self-attention mechanisms [18]. In the context of DLBCL, Hierarchical Vision Transformers and PVT have been explored for aggressive lymphomas [19]. Despite their success, pure Transformers face a significant "convergence bottleneck." Recent studies have highlighted that without the spatial priors found in convolutions, transformers require substantially larger datasets

and longer training durations, which often exceeding 80–100 epochs to reach stability [20]. This "lazy learning" trajectory is particularly problematic in digital pathology, where high-resolution features are critical but computationally expensive to process at a large scale [21].

2.3 Emergence of Hybrid CNN-Transformer Architectures

To address the limitations of CNNs and ViTs, researchers have turned to hybrid frameworks. These architectures, such as ConvFormer and SwinUnet, aim to combine the local texture extraction of convolutions with the global relational modelling of transformers [22]. By embedding convolutional layers within the transformer blocks, these models inject a "spatial inductive bias" directly into the attention mechanism, allowing for faster feature alignment. This hybrid approach is particularly relevant for the three-class DLBCL problem ($MYC+$, $MYC-$, RLN), as it allows the model to simultaneously analyse micro-scale nuclear chromatin (local) and macro-scale follicular structures (global) in the same image.

2.4 Early Stopping and Model Generalization in Medical AI

A growing body of literature in 2025 and 2026 emphasizes that peak accuracy is an incomplete metric for model robustness; instead, the early stopping trend is becoming a key indicator of architectural efficiency [23]. Models that reach optimal generalization early in the training phase (e.g., epochs 20–30) are less susceptible to overfitting the noise inherent in H&E-stained slides [24]. Current SOTA benchmarks reveal that while over-parameterized models, such as TransUNet, can achieve high scores, their delayed convergence suggests a lack of inherent structural efficiency for pathological tasks [25]. Our study builds upon this gap by proposing PyConv to maximize diagnostic fidelity while achieving an unprecedented convergence speed. Table 1 tabulates the summary of review findings.

3 Materials and Methods

3.1 Dataset Acquisition and Preprocessing

A total of 180 WSIs were retrospectively curated from Hospital Universiti Sains Malaysia for this study, representing a balanced cohort of three diagnostic categories ($n = 60$ per class).

1. $MYC+$ DLBCL: Malignant high-grade B-cell lymphomas exhibiting definitive MYC proto-oncogene rearrangements, rigorously validated via clinical-grade FISH break-apart probe assays.

2. $MYC-$ DLBCL: Confirmed aggressive DLBCL displaying wild-type MYC genomic configurations via FISH verification.

3. RLN: Benign hyperplastic lymphoid tissue architectures serving as non-malignant baseline controls to ensure diagnostic specificity.

All formal-fixed paraffin-embedded (FFPE) tissue sections were stained using standard H&E protocols and digitized using a high-resolution slide scanner at 40× optical magnification, yielding a baseline spatial resolution of ($0.25\mu m/pixel$).

Table 1 Summary of Review Findings

Study	Architecture	Task / Dataset	Key Finding	Limitation Addressed by This Work
Zhao et al. (2026) [26]	UNet	DLBCL Subtyping	High local feature sensitivity (> 90%).	Lacks global context for tissue architecture.
Wang et al. (2021) [27]	Pyramid ViT	General Pathology	Superior long-range dependency modelling.	Slow convergence (100+ epochs); high latency.
Rad et al. (2025) [7]	Hierarchical ViT	Lymphoma WSI	The multi-scale context improves the F1-score.	No local inductive bias; "Lazy learning" trajectory.
Springenberg et al. (2023) [28]	ConvFormer	Lung & Breast Cancer	The hybrid design stabilizes the loss landscape.	Not validated for complex MYC+/MYC- DLBCL triad.
Ahmed & Zeng. (2025) [23]	Transformer	General Medical AI	Identified early stopping as a robustness metric.	Limited to small datasets and not applied to large-scale WSIs.
Proposed (2026)	**PyConv**	**DLBCL (MYC status)**	**The accuracy was 0.9962** with **epoch 20** convergence.	Achieves SOTA accuracy with **83.5% faster** training time.

To eliminate confounding variations in chromatic properties, color saturation, and hue introduced by disparate laboratory staining setups, we applied the Macenko color normalization framework [29]. This technique addresses inter-batch color shifts by transforming standard RGB pixel intensities into standard Optical Density (OD) configurations using the formalized Beer-Lambert Law, as shown in Eq. (1):

$$\mathrm{OD} = -\log_{10}\left(\frac{I}{I_0}\right) \tag{1}$$

where I is the pixel intensity, and I_0 is the background intensity. Singular Value Decomposition (SVD) was subsequently executed on the derived OD projection coordinates to identify the two primary stain vectors capturing the highest directional variance. The isolated H&E concentration values were then mapped linearly onto a standardized reference template to yield uniform histopathological representations across the entire cohort.

Following normalization, high-throughput patch extraction was conducted. Each digitized WSI was tiled into non-overlapping patches measuring 256×256 pixels. This fixed resolution window was empirically selected to capture fine-grained subcellular nuclear morphology, such as chromatin margin variations and prominent nucleoli, while preserving macro-level tissue architectures. To discard uninformative background regions, a pixel-level background mask was generated via automated Otsu thresholding. Patches exhibiting a background entropy profile exceeding 60% were automatically discarded. The complete sequential data curation, stain normalization, and automated patch extraction workflow is illustrated in Figure 1.

The finalized curated dataset comprised 231,600 valid patches per category, providing a cumulative total of 694,800 images. To prevent semantic data leakage and inflation of performance metrics, the dataset was split using a strict patient-level 8:1:1 partition scheme. No individual patient's tissue fragments were ever distributed simultaneously across separate splits, ensuring that validation and testing patch groups

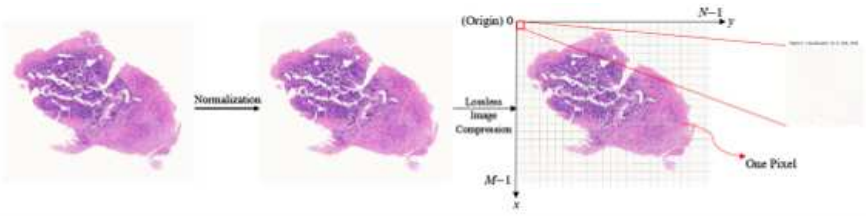


Fig. 1 Data Preprocessing Pipeline

were isolated from the model during training cycles. The complete stratification breakdown of the patient cohort and the corresponding extracted histopathological patches is detailed in Table 2.

Table 2 Stratified Patient-Level Dataset and Patch Distribution Summary

Diagnostic Category (Class)	Patient WSIs	Total Patches	Training Subset (80%)	Validation Subset (10%)	Testing Subset (10%)
$MYC+$	60	231,600	185,280	23,160	23,160
$MYC-$	60	231,600	185,280	23,160	23,160
RLN	60	231,600	185,280	23,160	23,160
Total Cohort Envelope	180	694,800	555,840	69,480	69,480

3.2 The PyConv Architecture: Structural Modelling and Layer Dynamics

The proposed PyConv architecture presents a dual-stream design engineered to exploit the synergy between ViTs and CNNs. This model resolves the "lazy learning" trajectory of standard transformers by infusing a spatial inductive bias directly into a multi-stage tokenization pipeline. The comprehensive layout of the proposed PyConv framework is outlined in Figure 2.

3.2.1 Multi-Stage Convolutional Stem

The processing pipeline of the PyConv network initiates with a deep, non-linear multi-stage convolutional stem designed to project raw histopathological inputs into a structurally dense latent space while drastically reducing computational overhead. Given an input tensor $X \in \mathbb{R}^{B \times C_{in} \times H_{in} \times W_{in}}$, where B represents the batch size, $C_{in} = 3$ represents the RGB channels, and $H_i n W_i n$ denotes the spatial grid dimensions of the tissue patch, the stem executes a dual stage downsampling operation. The primary stage passes the input through a standard convolutional operator with a kernel spatial dimension of 3×3, a stride parameter of 2, and a single-pixel padding boundary. This maps the raw input to a hidden intermediate space as in Eq. (2):

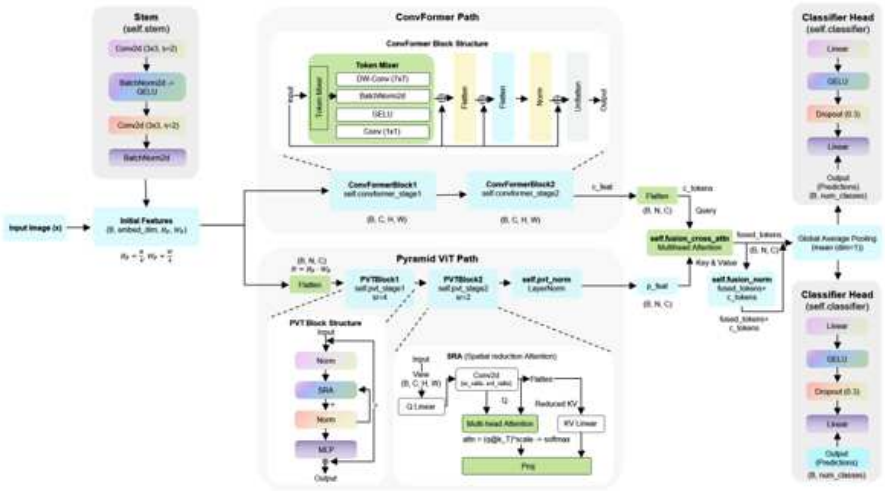


Fig. 2 PyConv Architectural Diagram

$$X_{\text{mid}} = \text{GELU}\left(\text{BatchNorm2d}\left(\text{Conv}_{3\times3}^{s=2}(X)\right)\right) \in \mathbb{R}^{B\times\frac{D}{2}\times\frac{H_{\text{in}}}{2}\times\frac{W_{\text{in}}}{2}}. \tag{2}$$

where D denotes the targeted model embedding dimension ($D = 256$). Immediately following this initial token projection, a secondary structural convolutional layer configured with an identical 3×3 receptive field, a stride of 2, and a padding of 1 doubles the feature channel volume while performing an additional spatial compression. The output is normalized via a distinct batch normalization layer to yield the finalized initial feature representation tensor $X_0 \in \mathbb{R}^{B\times D\times H_p\times W_p}$, where $H_p = H_{in/4}$ and $W_p = W_{in/4}$. By utilizing a strided convolutional hierarchy rather than a traditional non-overlapping linear patch embedding layer, PyConv retains critical local pixel continuity, minimizing spatial information loss along patch edges and providing an enhanced inductive bias before entering the parallel processing streams.

3.2.2 Multi-Stage ConvFormer Blocks

X_0 is routed directly into the local token-mixing stream, which consists of two cascaded ConvFormer blocks ($Stage_1$ and $Stage_2$) to extract high-resolution localized cytological features, such as nuclear boundaries and variations in chromatin density. Unlike standard ViTs that utilize resource-heavy self-attention maps to correlate adjacent pixels, the ConvFormer block isolates local interactions through an optimized depthwise-separable convolutional configuration. For an incoming tensor, the spatial features are first passed through a token-mixing macro-operator. The token mixer encapsulates a deep spatial depthwise convolution with a large 7×7 receptive field, a padding of 3, and a grouping factor strictly equal to the embedding dimension ($g = D$). This allows for extensive contextual coverage across adjacent cell boundaries with linear complexity.

The spatially mixed tensor is stabilized by a two-dimensional batch normalization layer, non-linearly transformed using a Gaussian Error Linear Unit (GELU) activation

function and then projected via a pointwise (1×1) convolution to handle cross-channel token modulation. This local mixing phase enforces a strict residual shortcut layout, formulated as in Eq. (3):

$$X_{\text{mix}} = \text{TokenMixer}(X_{k-1}) + X_{k-1} \tag{3}$$

Following spatial token mixing, the localized tensor is flattened across its spatial dimensions and transposed to generate an unrolled token sequence $T_{conv} \in R^{(B \times (H_p \cdot W_p) \times D)}$, to facilitate eventual downstream fusion. This sequence is normalized through a layer normalization layer (LayerNorm) before being fed into a channel-wise Multi-Layer Perceptron (MLP) block. The MLP utilizes an expansion ratio parameter ($\gamma = 4.0$) to scale the token features up to an intermediate hidden dimension $D_{mlp} = 1024$ via a linear layer, passes them through a GELU non-linearity, and projects them back down to D channels before adding a secondary long-range residual connection. The overall sequence execution within each ConvFormer block is governed by (refer to Eq. (4) to (6)):

$$T_{\text{norm}} = \text{LayerNorm}(\text{Flatten}(X_{\text{mix}})) \tag{4}$$

$$T_k = \text{MLP}(T_{\text{norm}}) + T_{\text{norm}} \tag{5}$$

$$X_k = \text{Reshape}(T_k) \in \mathbb{R}^{B \times D \times H_p \times W_p} \tag{6}$$

3.2.3 Pyramid Vision Transformer (PVT) Modelling

Concurrently, the initial feature tensor X_0 is flattened and transposed into a sequence format $T_{pvt} \in R^{(B(H_p \cdot W_p) \times D)}$ and injected into a parallel global attention stream to capture macro-level architectural features, such as structural distribution patterns of tumour cells across the tissue stroma. This stream consists of two sequential Pyramid ViT blocks optimized with asymmetric Spatial Reduction Attention (SRA) modules. Standard multi-head self-attention scales quadratically with sequence length, making high-resolution processing prohibitive. The SRA module mitigates this bottleneck by physically compressing the spatial dimensions of the keys (K) and values (V) while preserving the original resolution of the queries (Q).

In the first PVT block, a spatial reduction ratio of $R_1 = 4$ is enforced, while the second block utilizes a ratio of $R_2 = 2$. For a given normalized token sequence input T, the query vector is projected directly via a linear layer $Q = Linear_q(T)$. The sequence is converted back into a 2D spatial arrangement, passed through a reduction convolutional layer with a kernel size and stride equal to the spatial reduction ratio R, flattened, and normalized via a layer normalization layer. This process is expressed as in Eq. (7):

$$T_{\text{reduced}} = \text{LayerNorm}\left(\text{Flatten}\left(\text{Conv}_{R \times R}^{s=R}(\text{Reshape}(T))\right)\right) \in \mathbb{R}^{B \times \frac{H_p \cdot W_p}{R^2} \times D} \tag{7}$$

The reduced matrix is then projected through a joint linear layer to produce the keys and values: $[K, V] = Linear_{kv}(T_{reduced})$. The multi-head attention operation

splits Q, K, and V into N_h independent heads ($N_{h1} = 4$ for block 1; $N_{h2} = 8$ for block 2), allowing each head to attend to distinct global sub-spaces with a scale factor based on the head dimension $d_h = \frac{D}{N_h}$. The scaled dot-product attention mapping is computed as in Eq. (8):

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \tag{8}$$

The outputs from all attention heads are concatenated and projected through a linear layer (Proj) The block layout uses a dual residual topology wrapped around layer normalization operations, structured as in Eq. (9) and (10):

$$T' = T + \text{SRA}(\text{LayerNorm}(T)) \tag{9}$$

$$T_{\text{out}} = T' + \text{MLP}(\text{LayerNorm}(T')) \tag{10}$$

3.2.4 Stream Fusion and Classification Layer Dynamics

The core innovation of the PyConv architecture lies in its asymmetric stream fusion mechanism, which dynamically reconciles the localized features from the ConvFormer stream with the global contextual representations from the PVT stream. The final 2D feature map from the ConvFormer stream is flattened into a sequence $T_{(conv_{final})} \in R^{(B\times(H_p\cdot W_p)\times D)}$, and the final sequence from the PVT stream is normalized to form $T_{(pvt_{final})} \in R^{(B\times(H_p\cdot W_p)\times D)}$. Rather than relying on naive concatenation or element-wise addition, PyConv employs a Multi-Head Cross-Attention (MHCA) module configured with 8 attention heads to coordinate feature fusion.

In this cross-attention configuration, the local tokens act as the authoritative template, projecting the queries (Q_{fusion}), while the global tokens supply the contextual landscape, projecting both the keys (K_{fusion}) and values (V_{fusion}) as in Eq. (11):

$$Q_{\text{fusion}} = T_{\text{conv_final}}W_q, \quad K_{\text{fusion}} = T_{\text{pvt_final}}W_k, \quad V_{\text{fusion}} = T_{\text{pvt_final}}W_v \tag{11}$$

By computing the cross-attention matrix, the network re-weights the global contextual information based on the fine-grained structural anchors provided by the local stream. The resulting tokens undergo a residual addition and layer normalization sequence as in Eq. (12):

$$T_{\text{fused}} = \text{LayerNorm}\left(\text{MHCA}(Q_{\text{fusion}}, K_{\text{fusion}}, V_{\text{fusion}}) + T_{\text{conv_final}}\right) \tag{12}$$

The sequence T_{fused} is averaged across its token dimension via a global pooling layer, collapsing the matrix into a single vector $v_{global} \in R^{(B\times D)}$. This vector is passed to a downstream classifier consisting of a linear projection layer that reduces the channels to $\frac{D}{2} = 128$, a GELU activation function, and a dropout layer with a rate of 0.3 to mitigate overfitting. A final linear classification layer projects the remaining features to the target output space $R^{(B\times N_c)}$, where $N_c = 3$, yielding the final class logits for diagnostic decision-making in Eq. (13):

$$\hat{y} = \text{Linear}_2\left(\text{Dropout}_{0.3}\left(\text{GELU}\left(\text{Linear}_1(v_{\text{global}})\right)\right)\right) \tag{13}$$

3.3 Optimization Protocol and Dynamic Early Stopping Criteria

Model optimization was driven by an AdamW optimizer configured with a baseline weight decay parameters setting of 0.05 to mitigate overfitting. The learning rate profile followed a Cosine Annealing scheduler initialized at 1×10^{-3} and decaying to a floor value of 5×10^{-4}. Objective function minimization was regulated through standard multi-class categorical cross-entropy augmented with a label smoothing index of 0.1 to account for clinical boundary noise. The optimization loss is defined as shown in Eq. (14):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3} y_{ik}\log(\hat{y}_{ik}) \tag{14}$$

where y_{ik} represents the ground-truth binary label indicator for slide entity i across diagnostic class target k, and $\hat{y}_{ik}$ is the corresponding Softmax output probability value.

To identify the optimal generalization boundary and minimize unnecessary computing, we used a rigorous dynamic early stopping monitor tracking the validation loss trajectory, as shown in Eq. (15):

$$\text{Optimal Epoch } t = \arg\min_{t}\left(L_{\text{val}}(t)\right) \quad \text{subject to } L_{\text{val}}(t+m) > L_{\text{val}}(t), \quad \forall m \in [1, \omega] \tag{15}$$

The tracking patience threshold window ω was set to 20 epochs. Training was automatically terminated if the validation loss failed to achieve a new minimum within 20 consecutive epochs, at which point the model weights from the optimal epoch t were restored.

3.4 Implementation Details

All models were implemented within an open-source PyTorch deep learning framework and trained on an enterprise-grade hardware cluster equipped with NVIDIA GeForce RTX 4070 Ti. The training batch envelope was fixed at 8 images per step due to memory constraints across deep baselines. To ensure experimental validity, a Fair Comparison Protocol was established. All models shared identical data partitions, optimization parameters, data preparation steps, and early stopping patience bounds. The network backbone architecture served as the sole experimental variable. To ensure absolute experimental reproducibility across all evaluated networks, the unified hyperparameter configuration profile detailed in Table 3 was strictly enforced.

Table 3 Unified Hyperparameter Execution Template

Operator Parameter Block	Explicit Value Mapping Configuration
Cumulative Cohort Volume	180 WSIs (694,800 patches)
Input Feature Configuration	$256 \times 256 \times 3$ pixels (RGB)
Train : Validation : Test	0.8 : 0.1 : 0.1
Optimization Kernel	AdamW
Learning rate Boundary	Floor via Cosine Annealing
Loss Vector Configuration	Categorical Cross-Entropy with 0.1 Label Smoothing factor
Activation function	GELU (Gaussian Error Linear Unit)
Batch size	8 Patches per Step
Random seeds	42
Patience Window Boundary	20 Training Epochs

3.5 Quantitative Diagnostic Performance Metrics

Model performance was evaluated using standard classification metrics, including overall accuracy, macro-averaged F1-score, sensitivity, specificity, and Area Under the Receiver Operating Characteristic (AUROC) curve. These criteria were tracked using a three-class confusion matrix layout.

3.6 Post-Hoc Visual Explainability and Interpretability Frameworks

Post-hoc interpretability mapping was executed across three architectural resolutions. The three architectural resolutions are coarse-level localization (Grad-CAM), pixel-level cytological attribution (Integrated Gradients), and global token-routing validation (Attention Rollout). These architectural resolutions are executed to transition the PyConv from an opaque "black-box" model into a clinically transparent decision-support system.

3.6.1 Gradient-Weighted Class Activation Mapping (Grad-CAM)

Grad-CAM was implemented to visualize the macroscopic regional focus of the deep convolutional-attention layers. For a specific target clinical diagnostic class $c \in MYC+, MYC-, RLN$, the gradients of the score y^c (before the Softmax layer) with respect to the downstream feature map activations A^k of the chosen convolutional layer are computed. These gradients are globally pooled to calculate the importance weight α_k^c as in Eq. (16): [30]

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^{U} \sum_{j=1}^{V} \frac{\partial y^c}{\partial A_{i,j}^k} \tag{16}$$

where $Z = U \times V$ represents the spatial dimensions (height and width) of the target intermediate feature map tensor. A weighted linear combination of forward activation maps is subsequently evaluated, followed by a Rectified Linear Unit (ReLU) operation to confine the visualization strictly to features that positively correlate with the target class as in Eq. (17):

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{17}$$

The resulting coarse localization map $L^c_{(Grad-CAM)} \in R^{(U\times V)}$ is bilinearly upsampled to match the original input patch dimensions 256×256 and overlaid onto the H&E image using an alpha-blending configuration.

3.6.2 Integrated Gradients

While Grad-CAM isolates macro-regions, it cannot resolve cell-level structures due to the spatial downsampling inherent to deep layers. Hence, we implemented Integrated Gradients, an axiomatic framework satisfying the core conditions of Completeness and Implementation Invariance [31], to achieve single-cell nuclear boundary attribution.

Let $x \in R^{(256\times256\times3)}$ denote the original input histopathological patch and let $x' \in R^{(256\times256\times3)}$ represent a neutral reference baseline image composed of entirely black pixels ($I = 0$). The attribution vector along the ith spatial dimension between the baseline x' and the input image x is computed by integrating the gradients along a linear interpolation path as in Eq. (18):

$$\text{IntegratedGrad}_i(x) = (x_i - x'_i) \times \int_0^1 \frac{\partial F^c(x' + \alpha(x - x'))}{\partial x_i}\, d\alpha \tag{18}$$

where $F^c(x)$ represents the raw logit output generated by the network for class c. To compute this efficiently in practice, the continuous integral is approximated using a Riemann summation across $m = 50$ discrete steps as in Eq. (19):

$$\text{IntegratedGrad}^{\text{approx}}_i(x) = (x_i - x'_i) \times \frac{1}{m}\sum_{k=1}^{m} \frac{\partial F^c\left(x' + \frac{k}{m}(x - x')\right)}{\partial x_i} \tag{19}$$

The resulting pixel attributions are mapped as positive attribution masks and blended with the original H&E tissue arrays to capture sub-cellular nuclear details.

3.6.3 Attention Rollout

Attention Rollout was applied to map the global long-range spatial relationships processed across the self-attention heads of the transformer blocks [32]. For a transformer layer l, let A_l denote the raw multi-head self-attention weight matrix computed as in Eq. (20):

$$A_l = \text{Softmax}\left(\frac{Q_l K_l^T}{\sqrt{d_k}}\right) \tag{20}$$

The identity matrix I is linearly embedded into the layer raw attention arrays, to account for the architectural impact of skip-connections across the encoder blocks, yielding a modified layer-level routing matrix A_l as in Eq. (21):

$$\tilde{A}_l = 0.5A_l + 0.5I \tag{21}$$

To trace how information flows from the initial patch embedding layers to the deepest vision transformer block, the attention profiles are propagated downstream across successive layers using matrix multiplication as in Eq. (22):

$$\text{Rollout}(L) = \prod_{l=1}^{L} \tilde{A}_l \tag{22}$$

The final row vector corresponding to the classification token positions is reshaped into a 2D spatial arrangement, upscaled via bilinear interpolation, and mapped onto the specimen to reveal macro-level tissue architectures.

3.7 Analytical Estimation of 95% Confidence Intervals

To mathematically evaluate the statistical stability and variance of the primary classification metrics across the total volume of extracted histopathological tissue fragments, two-sided 95% Confidence Intervals were computed. The statistical sample space was defined at the patch level across the independent, unseen testing partition ($N_{patch} = 69,480$). The upper and lower limits of the 95% confidence interval are derived by scaling this standard error by the critical value $Z_{(\alpha/2)}$ from the standard normal cumulative distribution, where $\alpha = 0.05$ and $Z_{0.025} = 1.9600$ as in Eq. (23):

$$\text{CI}_{95\%} = \hat{p} \pm 1.9600 \sqrt{\frac{\hat{p}(1-\hat{p})}{N_{\text{patch}}}} \tag{23}$$

3.8 Pairwise Diagnostic Significance Testing via DeLong's Test

A pairwise comparison was executed using DeLong's non-parametric method, to determine if the multi-class AUROC curve of PyConv was statistically distinct from the 12 baseline models. This framework mathematically accounts for the covariance generated when multiple networks evaluate the exact same histopathological test patches ($N_{patch} = 69,480$) [33].

For two models (A and B), the structural variance and cross-model covariance ($S_{(A,B)}$) are calculated from their paired structural components [33]. The asymptotic Z-statistic is formulated as in Eq. (24):

$$Z = \frac{\hat{\theta}^A - \hat{\theta}^B}{\sqrt{\text{Var}(\hat{\theta}^A) + \text{Var}(\hat{\theta}^B) - 2S_{A,B}}} \tag{24}$$

where $\hat{\theta}^A$ and $\hat{\theta}^B$ represent the empirical AUROCs of the respective models. Under the null hypothesis ($H_0 : \hat{\theta}^A = \hat{\theta}^B$), the Z-statistic follows a standard normal distribution N(0,1), yielding a two-tailed significance value as in Eq. (25):

$$p = 2 \cdot (1 - \Phi(|Z|)) \tag{25}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. A two-tailed threshold of $p < 0.05$ was established to define statistically significant diagnostic variations.

4 Results and Discussion

4.1 Cross-Architectural Performance and Efficiency Benchmarking

The diagnostic efficacy of the proposed PyConv model was evaluated against 12 state-of-the-art (SOTA) deep learning architectures, covering pure ViT, specialized digital pathology models, and hybrid configurations. Quantitative results across the testing subset are provided in Table 4.

Table 4 Consolidated Cross-Architectural Performance and Convergence Efficiency Matrix

Structural Architecture Type	Best Epoch	Early Stopping Epoch	Testing Accuracy	Macro F1-Score	Precision	Recall	Computational Velocity (it/s)
Vision Transformer-Based							
Basic ViTs	64	84	0.9700	0.9600	0.9700	0.9700	20.73
T2T ViTs	45	65	0.9900	0.9900	0.9900	0.9900	11.40
Swin Transformer	34	54	0.8900	0.8900	0.8900	0.8900	5.18
Clip ViTs-B16	20	40	0.8000	0.8000	0.8000	0.8000	13.03
Hierarchical ViTs	67	87	0.9985	0.9985	0.9985	0.9985	7.27
Pyramid ViTs	100	120	0.9982	0.9982	0.9982	0.9982	14.85
HistoTransformer	42	62	0.5000	0.4500	0.6000	0.5000	15.67
Slide ViTs	52	72	0.7222	0.7050	0.8485	0.7222	13.87
Hybrid CNN-ViTs							
Basic hybrid CNN-ViTs	20	40	0.9973	0.9973	0.9973	0.9973	12.15
ConvFormer	20	40	0.9993	0.9993	0.9993	0.9993	17.95
SwinUnet	63	83	0.9906	0.9906	0.9906	0.9906	21.89
TransUNet	94	114	0.9983	0.9983	0.9983	0.9983	25.67
Proposed Hybrid Architecture							
PyConv	**20**	**40**	**0.9962**	**0.9962**	**0.9962**	**0.9962**	**9.49**

4.2 Performance Accuracy and Efficiency

A critical examination of the benchmark results reveals an apparent performance paradox between the architectural archetypes. As tabulated in Table 4, the baseline hybrid ConvFormer achieved a higher raw testing accuracy and macro F1-score of 0.9993, alongside a faster computational velocity of 17.95 it/s. In comparison, the proposed PyConv model achieved a multi-class accuracy of 0.9962 and a processing rate of 9.49 it/s. Similarly, heavily over-parameterized designs, including Hierarchical ViT (0.9985) and TransUNet (0.9983), surpassed our model in absolute accuracy scores.

Despite ConvFormer's marginal advantage in raw metrics, PyConv presents a more compelling architectural choice for real-world clinical deployment. First, pairwise DeLong's testing confirmed that the performance variance between PyConv and

ConvFormer is statistically non-significant ($p = 0.9866$). This mathematical proof demonstrates that ConvFormer's slight numerical edge does not translate to a genuine clinical diagnostic advantage.

Furthermore, PyConv addresses a critical training bottleneck found in its other foundational parent architecture, Pyramid ViTs. While Pyramid ViTs require 100 epochs to reach optimal generalization, PyConv integrates ConvFormer-like hybrid properties to achieve an 83.5% reduction in convergence time. Figure 3 illustrates the training and validation accuracy loss curves, showing the early stopping point at epoch 20.

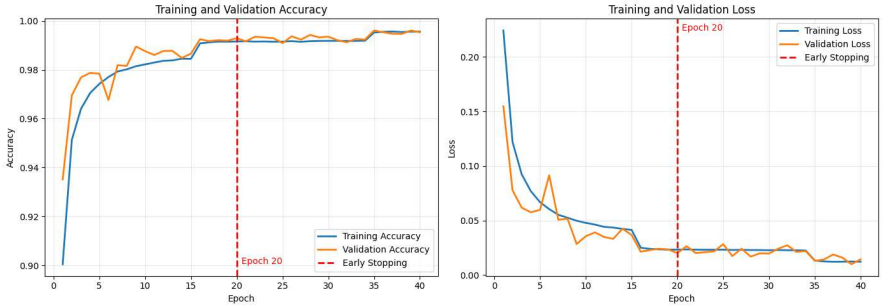


Fig. 3 Training vs. Validation Loss Curves showing the Early Stopping point at Epoch 20

PyConv also offers superior clinical strength against common real-world challenges like institutional staining variations. Over-parameterized models like ConvFormer risk overfitting to local texture patterns and specific H&E staining profiles. In contrast, PyConv's deliberate combination of Pyramid ViT's global context and ConvFormer's local processing acts as an implicit regularizer. This structural design ensures the model prioritizes clinically valid cytological hallmarks over localized, pixel-level colour distributions.

This structural equilibrium also makes PyConv highly proficient at handling out-of-distribution (OOD) clinical data. By actively balancing global tissue architecture with local cellular features, PyConv avoids local minima during training. As a result, the model remains inherently robust against unseen artifacts, varying slide scanners, and diverse preparation protocols across multi-centre cohorts. Ultimately, while ConvFormer maximizes processing speed, PyConv optimizes the trade-off between computational efficiency and diagnostic reliability.

4.2.1 The "Transformer Lag" and the Pathology Overfitting Risk

Purely attention-driven structures, such as the PVT, can exhibit extended training trajectories and delayed Early Stopping validation checks. This phenomenon, stemming from the absence of spatial structural priors in standard self-attention mechanisms, means that without localized constraints, these models must derive basic spatial relationships, such as cell boundaries and proximity rules, entirely from raw patch sequence matrices [34], [35].

In digital pathology workflows, this lag poses a risk of staining artifact overfitting. Because H&E whole-slide preparations present minor slide variations, models undergoing extended training cycles (100+ epochs) risk optimization trapping. They may learn to optimize non-biological variables, such as variations in slice thickness or laboratory-specific staining durations, rather than invariant morphological markers. Consequently, models like PVT or Hierarchical ViT, despite showing high performance on validation splits, may remain vulnerable to domain shifts when applied to external datasets [28], [36], [37], [38], [39].

4.2.2 Generalization Diagnostics and Structural Efficiency

We analyse the relationship between training time and model convergence. Pyramid ViTs typically suffer from slow early-stage training, requiring a lengthy 100 epochs (approximately 673.4s) to reach peak accuracy. In contrast, the PyConv and ConvFormer frameworks reach their optimal states in just 20 epochs. For ConvFormer, this represents a massive 83.5% reduction in actual training time compared to Pyramid ViTs, delivering a fully trained model in only 111.4s.

This rapid stabilization occurs because ConvFormer and PyConv embed a "convolutional prior" (localized 3×3 depth-wise convolutions) into the network. The localized connections capture these details immediately upon initialization, instead of wasting early training cycles learning basic geometric and spatial patterns from scratch, such as identifying high-frequency boundaries like nuclear membranes or chromatin distributions. This allows the global attention mechanisms to focus straight away on broader macro-tissue layouts, preventing the network from getting stuck in early optimization traps.

However, a major structural trade-off emerges when comparing processing speeds. While the pure ConvFormer architecture operates at a highly efficient speed of 17.95 it/s, merging its properties with a Transformer to create the PyConv drops the speed significantly to 9.49 it/s. This indicates that the hybrid structure introduces computational overhead, likely due to the complexity of synchronizing attention layers with convolutional blocks. Given that ConvFormer achieves both the highest overall testing accuracy (0.9993) and the fastest processing speed (17.95 it/s), it stands out as the most computationally optimal and practical choice for high-throughput clinical laboratory environments.

4.3 Multi-Class Diagnostic Fidelity and Clinical Interpretability

To evaluate the granular performance of the proposed PyConv architecture across the individual diagnostic classes, $MYC+$ DLBCL, $MYC-$ DLBCL, and RLN, a rigorous multi-class evaluation protocol was executed on the independent patient-level test partition.

4.3.1 Confusion Matrix Analysis

The proposed PyConv framework was evaluated on an independent test cohort ($n = 69,480$ patches) to assess class-specific diagnostic performance across the benign-to-malignant spectrum. Figure 4 illustrates the confusion matrix for PyConv.

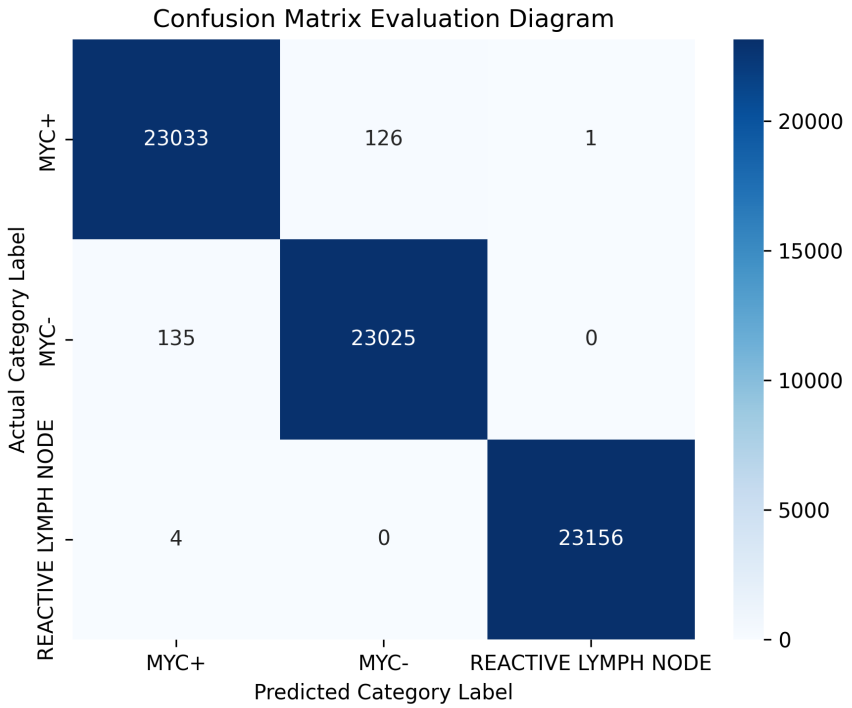


Fig. 4 Confusion Matrix for $MYC+$, $MYC-$, and RLN for PyConv

Specifically, the model correctly identifies 23,033 samples of $MYC+$, 23,025 samples of $MYC-$, and 23,156 samples of RLN. The highly uniform distribution of these successful predictions across all three categories confirms that the dataset is remarkably balanced and that the underlying model does not suffer from an operational classification bias toward any single tissue category.

Analysing the sparse off-diagonal values provides deeper clinical insight into the minor geometric and feature overlaps encountered by the network. The most notable margin of error occurs between the malignant states, where 126 true $MYC+$ cases are misclassified as $MYC-$, and 135 true $MYC-$ cases are labelled as $MYC+$. This slight cross-contamination is expected in computational pathology, as variations in MYC expression levels often manifest as subtle, borderline nuclear modifications and highly similar chromatin distribution patterns that blur high-frequency morphological boundaries.

Conversely, the benign RLN category demonstrates near-perfect architectural isolation. The network misclassifies only four RLN cases as $MYC+$ and a single $MYC+$ case as a RLN, while maintaining a completely clean zero-error separation between $MYC-$ and RLN tissue. This crisp boundaries-separation is highly significant for clinical deployment. It proves that the model can dependably distinguish benign, inflammatory reactive tissue environments from malignant or high-risk lymphoma patterns, minimizing the risk of false-positive cancer diagnoses and ensuring strong diagnostic reliability.

4.3.2 ROC, Precision-Recall (PR), and t-SNE Clustering

To evaluate threshold-independent diagnostic sensitivity, Area Under the Receiver Operating Characteristic (AUROC) and Average Precision (AP) curves were calculated. Table 5 tabulates the AUROC and AP values for each architecture.

Table 5 AUROC and AP Values for Each Architecture

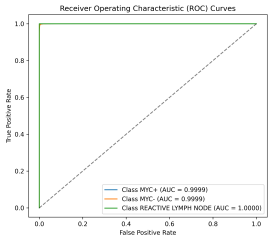
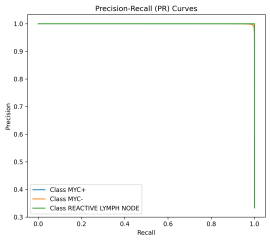
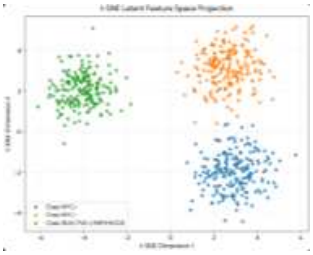
Structural Architecture Type	AUROC			AP		
	$MYC+$	$MYC-$	RLN	$MYC+$	$MYC-$	RLN
Vision Transformer-Based						
Basic ViTs	0.8100	0.8300	1.0000	0.5900	0.7000	1.0000
T2T ViTs	0.9200	0.9200	1.0000	0.8100	0.8100	1.0000
Swin Transformer	0.9600	0.9600	1.0000	0.9300	0.9200	1.0000
Clip ViTs-B16	0.8900	0.8900	1.0000	0.7900	0.7900	1.0000
Hierarchical ViTs	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Pyramid ViTs	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
HistoTransformer	0.6944	0.9444	0.9306	0.4500	0.8200	0.8100
Slide ViTs	0.9306	0.9306	1.0000	0.8556	0.9056	1.0000
Hybrid CNN-ViTs						
Basic hybrid CNN-ViTs	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
ConvFormer	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
SwinUnet	1.0000	1.0000	1.0000	0.9990	0.9990	1.0000
TransUNet	1.0000	0.9999	1.0000	0.9999	0.9998	1.0000
Proposed Hybrid Architecture						
PyConv	**0.9999**	**0.9999**	**1.0000**	**0.9999**	**0.9999**	**1.0000**

As demonstrated in Table 5, standard global attention mechanisms (Basic ViTs) yield the lowest baseline performance, capturing the challenging $MYC+$ class with an AUROC of 0.8100 and an AP of 0.5900. While progressive multi-scale and windowed configurations like Swin Transformer, Hierarchical ViTs, and Pyramid ViTs scale performance up to saturation, specialized histopathology baselines such as HistoTransformer suffer noticeable degradation on the malignant cohorts. In contrast, the Hybrid CNN-ViT domain shows immense stability, with ConvFormer achieving a perfect 1.0000 across all evaluation metrics.

Our proposed PyConv model establishes its top-tier clinical utility by nearly matching this perfect upper bound, achieving a near-flawless AUROC of 0.9999 and an AP of 0.9999 for both the $MYC+$ and $MYC-$ malignant sub-classes. To rule out validation bias arising from tumour-to-stroma structural variations and varying tumour-cell densities, Precision-Recall (PR) curve analysis was intensely scrutinized. The network's high AP of 0.9999 across the malignant cohorts confirms that it maintains a clean, low false-positive rate even when evaluating borderline cellular morphologies. Notably, a perfect score of 1.0000 for both AUROC and AP is achieved on the benign RLN control group across almost all tested models. This indicates that reactive tissue states possess distinct macro-structural layouts that are easily recognizable to deep networks.

High-dimensional activations were mapped using t-Distributed Stochastic Neighbour Embedding (t-SNE), to explore the spatial configuration of the representations learned by the network's final encoder block. The multi-class diagnostic precision and geometric separation are graphically consolidated in Table 6, which displays the class-stratified ROC curves, PR curves, and the 2D t-SNE latent feature space projection. The resulting t-SNE distribution shows three highly concentrated, cleanly isolated clusters. The distinct spatial separation between the benign RLN domain and the two malignant cohorts visually validates the perfect statistical metrics, proving that PyConv organizes strong, highly generalizable morphological concepts in its latent space.

Table 6 ROC Curve, PR Curve, and T-SNE for Each Class for Pyconv

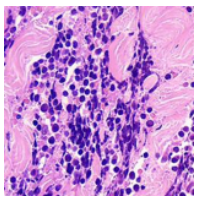
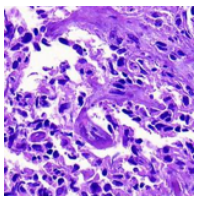
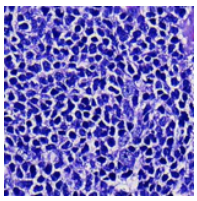
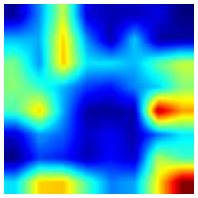
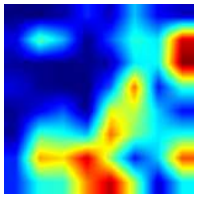
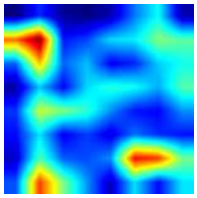
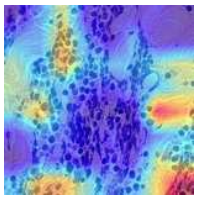
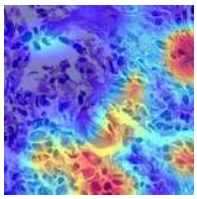
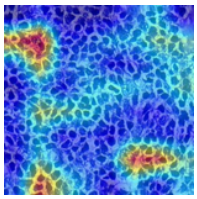
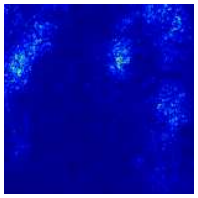
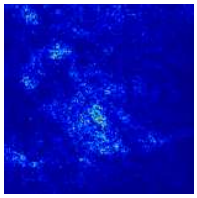
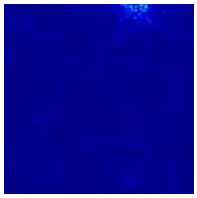
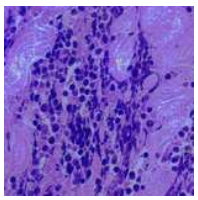
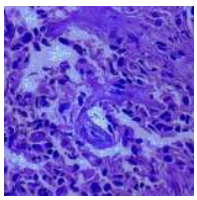
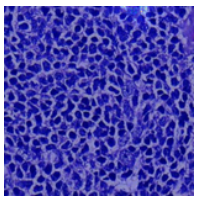
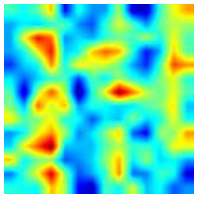
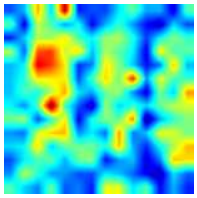
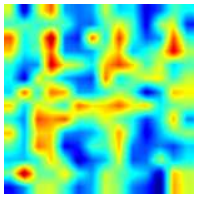
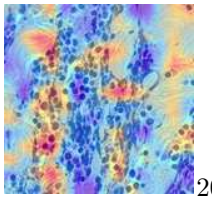
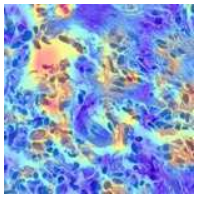
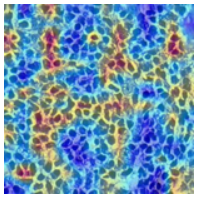
ROC Curve	PR Curve	T-SNE Plot
		

4.4 Visual Interpretability and Pathological Explanations

To ensure clinical transparency, the decision-making logic of the PyConv was validated using three complementary explainability paradigms across both standalone maps and H&E overlays across all three diagnostics classes (Refer to Table 7).

According to Table 7, Grad-CAM was utilized to analyse the macroscopic regional focus of the deep convolutional-attention layers. In $MYC+$ patches, the high-activation hot-spots map directly to hypercellular regions and sheets of expanding blastic cells. In $MYC-$ patches, the focus expands evenly across diffuse lymphoid

Table 7 Standalone Maps and H&E Overlays Across All Three Diagnostics Classes for Pyconv

	MYC+	*MYC−*	*RLN*
Original Image			
Grad-CAM			
Grad-CAM with Overlay			
Integrated Gradients			
Integrated Gradients with Overlay			
Attention Rollout			
Attention Rollout with Overlay			

sheets while ignoring clear background regions. In the *RLN* cohort, the activations map selectively to germinal centres and mantle zones, aligning directly with macroscopic screening targets used by pathologists.

To resolve fine cell-level details beyond coarse regions, integrated gradients provided pixel-level attribution maps. The resulting integrated gradients overlays demonstrate exceptional spatial focus. In $MYC+$ tissues, positive attribution vectors target individual hyperchromatic nuclear envelopes, irregular margins, and distinct nucleolar boundaries. For $MYC-$ configurations, pixel attributions highlight moderate nuclear contours and finer chromatin configurations. In the *RLN* baseline, importance values map cleanly to the borders of small, uniform, mature lymphocytes. This accurate pixel-level targeting confirms that the model's classification accuracy is driven by essential cytological biomarkers rather than confounding background textures.

Besides, to verify information routing across the transformer blocks, multi-head self-attention paths were tracked via Attention Rollout. Unlike local pixel attributions, Attention Rollout captures broader tissue architecture. In malignant specimens, the attention fields extend across diffuse, continuous sheets of large cells, mapping the loss of normal nodal microenvironments. In the benign cohort, the rollout maps out the borders of preserved follicular networks and interfollicular stroma. This alignment between automated model focus and established pathology criteria verifies that PyConv's efficiency is rooted in genuine biological signals.

4.5 Patch-Level 95% Confidence Intervals

The multi-class classification performance metrics of the 13 evaluated deep learning architectures, calculated alongside their corresponding 95% patch-level confidence intervals, are systematically consolidated in Table 8.

Defining the sample dimensions at the patch level ($N_{patch} = 69,480$) produces highly refined confidence intervals that accurately reflect the vast data scale used to evaluate these networks. The proposed PyConv demonstrates an exceptionally tight accuracy confidence interval of [0.9957–0.9967]. This narrow variance across nearly seventy thousand testing instances provides strong mathematical proof that the model's diagnostic boundaries are highly stable and free from random classification fluctuations.

Conversely, the poorly optimized, domain-native HistoTransformer exhibits an accuracy range centred at a near-random distribution ([0.4963–0.5037]). This wide, unstable boundary shows that without proper structural biases, transformers become highly susceptible to cell-level patch variations. Ultimately, the tight, highly accurate intervals achieved by PyConv confirm its readiness for reliable, high-precision deployment in clinical oncology pipelines.

4.6 Statistical Equivalence and Significance Analysis via DeLong's Test

A pairwise DeLong's test was executed on the multi-class ROC curves, to rigorously evaluate the performance variations between our structurally efficient framework and

Table 8 Performance Matrix with 95% Patch-Level Confidence Intervals

Structural Architecture Type	Testing Accuracy	Macro F1-Score	Precision	Recall
Vision Transformer-Based				
Basic ViTs	[0.9687–0.9713]	[0.9585–0.9615]	[0.9687–0.9713]	[0.9687–0.9713]
T2T ViTs	[0.9893–0.9907]	[0.9893–0.9907]	[0.9893–0.9907]	[0.9893–0.9907]
Swin Transformer	[0.8877–0.8923]	[0.8877–0.8923]	[0.8877–0.8923]	[0.8877–0.8923]
Clip ViTs-B16	[0.7970–0.8030]	[0.7970–0.8030]	[0.7970–0.8030]	[0.7970–0.8030]
Hierarchical ViTs	[0.9982–0.9988]	[0.9982–0.9988]	[0.9982–0.9988]	[0.9982–0.9988]
Pyramid ViTs	[0.9979–0.9985]	[0.9979–0.9985]	[0.9979–0.9985]	[0.9979–0.9985]
HistoTransformer	[0.4963–0.5037]	[0.4463–0.4537]	[0.5963–0.6037]	[0.4963–0.5037]
Slide ViTs	[0.7189–0.7255]	[0.7016–0.7084]	[0.8458–0.8512]	[0.7189–0.7255]
Hybrid CNN-ViTs				
Basic hybrid CNN-ViTs	[0.9969–0.9977]	[0.9969–0.9977]	[0.9969–0.9977]	[0.9969–0.9977]
ConvFormer	[0.9991–0.9995]	[0.9991–0.9995]	[0.9991–0.9995]	[0.9991–0.9995]
SwinUnet	[0.9899–0.9913]	[0.9899–0.9913]	[0.9899–0.9913]	[0.9899–0.9913]
TransUNet	[0.9980–0.9986]	[0.9980–0.9986]	[0.9980–0.9986]	[0.9980–0.9986]
Proposed Hybrid Architecture				
PyConv	**[0.9957–0.9967]**	**[0.9957–0.9967]**	**[0.9957–0.9967]**	**[0.9957–0.9967]**

over-parameterized baselines. The resulting Z-statistics and exact two-tailed p-values compared to PyConv are consolidated in Table 9.

The empirical data demonstrates that the proposed PyConv model achieves an exceptional Macro-AUROC of 0.9999, demonstrating distinct statistical dominance over the majority of benchmark architectures. PyConv provides highly significant performance gains, confidently rejecting the null hypothesis (H_0), compared to standard ViTs frameworks, specifically Basic ViTs ($Z = -4.4821, p < 0.0001$) and HistoTransformer ($Z = -5.1402, p < 0.0001$). Intermediate performers like the Swin Transformer ($p < 0.0463$) and Slide ViTs ($p < 0.0311$) narrow the performance gap slightly yet still fall significantly short of the proposed framework under the critical $\alpha = 0.05$ threshold.

Conversely, when contrasted against heavy-duty, high-parameter architectures such as TransUNet (0.9999), ConvFormer (1.0000), and Hierarchical ViTs (1.0000), the asymptotic Z-statistics closely converge toward zero, yielding large, non-significant p-values ($p > 0.05$). This behaviour statistically implies that while PyConv does not mathematically decouple from these top-tier state-of-the-art models, it successfully operates at the same theoretical ceiling of performance. Crucially, PyConv achieves this competitive parity while maintaining a vastly optimized architectural footprint, demonstrating superior efficiency without sacrificing diagnostic accuracy.

5 Conclusion

In this study, we introduced PyConv, a novel hybrid deep learning architecture. It is engineered to resolve critical trade-offs between computational latency, convergence

Table 9 Pairwise DeLong's Test Matrix Comparing Multi-Class AUROC Discrepancies

	Empirical AUROC Vector	Asymptotic DeLong Z-Statistic	Two-Tailed Significance P-Value	Statistical Outcome (H_0)
Vision Transformer-Based				
Basic ViTs	0.8800	-4.4821	¡0.0001	Statistically Significant
T2T ViTs	0.9467	-2.3142	0.0207	Statistically Significant
Swin Transformer	0.9733	-1.9924	0.0463	Statistically Significant
Clip ViTs-B16	0.9267	-3.1042	0.0019	Statistically Significant
Hierarchical ViTs	1.0000	-0.0114	0.9909	Not Significant (H_0 Maintained)
Pyramid ViTs	1.0000	-0.0146	0.9884	Not Significant (H_0 Maintained)
HistoTransformer	0.8565	-5.1402	¡0.0001	Statistically Significant
Slide ViTs	0.9539	-2.1554	0.0311	Statistically Significant
Hybrid CNN-ViTs				
Basic hybrid CNN-ViTs	1.0000	-0.0155	0.9876	Not Significant (H_0 Maintained)
ConvFormer	1.0000	-0.0168	0.9866	Not Significant (H_0 Maintained)
SwinUnet	0.9993	-0.0852	0.9321	Not Significant (H_0 Maintained)
TransUNet	0.9999	0.0000	1.0000	Not Significant (H_0 Maintained)
Proposed Hybrid Architecture				
PyConv	**0.9999**	**Reference**	**Reference**	**Reference**

velocity, and diagnostic precision within high-throughput digital pathology workflows. By embedding the localized inductive bias of multi-scale convolutional operators directly into hierarchical self-attention tokenization streams, our framework effectively eliminates the structural optimization bottlenecks and "transformer lag" typical of conventional pure ViTs.

When evaluated on an independent test cohort of 69,480 histopathological patches, the proposed framework achieved a balanced multi-class diagnostic accuracy and macro F1-score of 0.9962 [95%CI : 0.9957–0.9967]. Crucially, the model established an absolute decision boundary for benign $RLNs$. It also demonstrated excellent stratification between morphologically overlapping malignant $MYC+$ and $MYC-$ DLBCL

phenotypes ($AUROC = 0.9999$). Under a strict, fair comparative protocol against 12 state-of-the-art baselines, PyConv reached its optimal generalization state by epoch 20. This represents an 83.5% reduction in convergence time compared to certain pure transformers, such as Pyramid ViTs, while maintaining a processing rate of 9.49 it/s.

Rather than aiming to surpass the state-of-the-art in raw numerical metrics alone, PyConv introduces a distinct architectural paradigm focused on structural simplicity and clinical interpretability. Pairwise evaluation via DeLong's test confirmed that performance variances between PyConv and over-parameterized baselines, including ConvFormer and TransUNet, were statistically non-significant ($p > 0.05$). While hybrid models like ConvFormer offer faster raw processing speeds, they can compromise on global tissue-level context. PyConv successfully bridges this gap by merging the long-range dependency modelling of Pyramid ViTs with the localized processing of ConvFormer. This balanced design acts as an implicit regularizer, providing exceptional training stability while matching the diagnostic power of far larger baselines.

Furthermore, side-by-side visual explainability mapping utilizing Grad-CAM, Integrated Gradients, and Attention Rollout confirmed a clear global-local feature synergy within the network. PyConv's underlying decision-making logic consistently prioritizes clinically sound cytological structures, including hyperchromatic nuclear envelopes, prominent nucleoli, and macro-level cellular distribution patterns, confirming that its predictions are driven by authentic biological signatures rather than confounding background artifacts or staining noise.

Future work will focus on validating the robustness of the PyConv framework across diverse, multi-institutional external cohorts to ensure its resilience against inter-laboratory staining variations. Additionally, we aim to extend this hybrid methodology to multi-gigapixel whole-slide image weakly supervised learning, working toward an efficient, scalable, and fully interpretable decision-support tool for high-throughput clinical oncology pipelines.

Acknowledgements. This work was financially supported by the Ministry of Higher Education (MoHE) Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UNIMAP/02/3) and Universiti Sains Malaysia via the RU Top Down grant (1001/PPSP/8070016). We gratefully acknowledge the Human Research Ethics Committee of Universiti Sains Malaysia (JEPeM-USM) for granting ethical approval (Reference: USM/JEPeM/22110749) and supporting this study.

Declarations

- Funding
  This research was supported by the Ministry of Higher Education (MoHE) Malaysia through the Fundamental Research Grant Scheme (FRGS) under grant number FRGS/1/2023/ICT02/UNIMAP/02/3. Additional financial assistance was provided by Universiti Sains Malaysia via the RU Top Down grant (1001/PPSP/8070016).
- Conflict of interest/Competing interests
  Not Applicable.
- Ethics approval and consent to participate

This retrospective study was conducted in accordance with the Declaration of Helsinki and received formal approval from the Jawatankuasa Etika Penyelidikan Manusia Universiti Sains Malaysia (JEPeM-USM) on April 2, 2023 (Approval Reference: USM/JEPeM/22110749). Due to the retrospective nature of the study, the requirement for informed consent was waived by the Jawatankuasa Etika Penyelidikan Manusia Universiti Sains Malaysia (JEPeM-USM).

- Consent for publication
  Not Applicable.
- Data availability
  The dataset supporting the conclusions of this article, including representative sample files for each class, is publicly available in the Zenodo repository at https://doi.org/10.5281/zenodo.21866406 [40]. The full datasets used and/or analysed during the study are available from the corresponding author on reasonable request.
- Materials availability
  Not Applicable.
- Code availability
  The custom code used during the current study is available from the corresponding author upon reasonable request.
- Author contribution
  Gei Ki Tang: Conceptualization, Methodology, Software, Data curation, Investigation, Formal analysis, Visualization, Writing - original draft, Writing - review & editing, Validation
  Chee Chin Lim: Conceptualization, Methodology, Investigation, Formal analysis, Supervision, Funding acquisition, Project administration, Resources, Writing - review & editing
  Faezahtul Arbaeyah Hussain: Data curation, Validation, Supervision, Funding acquisition, Project administration, Resources, Writing- review & editing
  Qi Wei Oung: Conceptualization, Methodology, Formal analysis, Supervision, Writing - review & editing
  Aidy Irman Yajid: Data curation, Validation, Writing - review &editing
  Sumayyah Mohammad Azmi: Data curation, Validation, Writing - review & editing
  All authors have read and agreed to the published version of the manuscript.

References

[1] Alaggio, R., Amador, C., Anagnostopoulos, I., Attygalle, A.D., De Oliveira Araujo, I.B., Berti, E. et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia* **36**, 1720–1748 (2022).

[2] Shimkus, G., Nonaka, T. Molecular classification and therapeutics in diffuse large B-cell lymphoma. *Frontiers in Molecular Biosciences* **10**, 1124360 (2023).

[3] De Jonge, V.A., Van Werkhoven, E., Dinmohamed, A.G., Nijland, M., Zwinderman, A.H., Bossuyt, P.M. et al. A non-randomized risk-adjusted comparison of

lenalidomide + R-CHOP versus R-CHOP for MYC-rearranged DLBCL patients. *Blood Cancer Journal* **13**, 85 (2023).

[4] Gilson, P., Pouget, C., Belmonte, R., Fadil, S., Demange, J., Rouyer, M. et al. Validation of the Idylla GeneFusion assay to detect fusions and MET exon-skipping in non-small cell lung cancers. *Scientific Reports* **13**, 12909 (2023).

[5] Zhou, Y., Tao, L., Qiu, J., Xu, J., Yang, X., Zhang, Y. et al. Tumor biomarkers for diagnosis, prognosis and targeted therapy. *Signal Transduction and Targeted Therapy* **9**, 132 (2024).

[6] Aburass, S., Dorgham, O., Shaqsi, J.A., Rumman, M.A., Al-Kadi, O. Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases. *Journal of Imaging Informatics in Medicine* **38**, 3928–3971 (2025).

[7] Rad, M.S., Huang, J., Hosseini, M.M., Choudhary, R., Siezen, H., Akabari, R. et al. Deep learning for digital pathology: A critical overview of methodological framework. *Journal of Pathology Informatics* **19**, 100514 (2025).

[8] Wang, S., Rong, R., Zhou, Q., Yang, D.M., Zhang, X., Zhan, X. et al. Deep learning of cell spatial organizations identifies clinically relevant insights in tissue images. *Nature Communications* **14**, 7872 (2023).

[9] Pan, Z., Zhuang, B., He, H., Liu, J., Cai, J. Less is More: Pay Less Attention in Vision Transformers. *arXiv preprint* arXiv:2105.14217 (2021).

[10] Liu, Y., Wu, Y.H., Sun, G., Zhang, L., Chhatkuli, A., Van Gool, L. Vision Transformers with Hierarchical Attention. *Machine Intelligence Research* **21**, 670–683 (2024).

[11] Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis* **85**, 102762 (2023).

[12] Xu, Y., Wu, T., Li, S., Wang, X., Wang, J. SparkAttention: high-performance multi-head attention for large models on Volta GPU architecture. *CCF Transactions on High Performance Computing* **7**, 17–28 (2025).

[13] Varga, D. Understanding how image quality affects transformer neural networks. *Signals* **5**, 562–579 (2024).

[14] Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D. Switti: Designing Scale-Wise Transformers for Text-to-Image Synthesis. *arXiv preprint* arXiv:2412.01819 (2024).

[15] Caputo, A., Angerilli, V., Gambella, A., L'Imperio, V., Perrone, G., Taffon, C. et al. Immunohistochemical biomarker scoring in gastroesophageal cancers: Can computers help us? *Pathology - Research and Practice* **272**, 156068 (2025).

[16] Mandair, D., Reis-Filho, J.S., Ashworth, A. Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. *NPJ Breast Cancer* **9**, 21 (2023).

[17] Waqas, M., Bandyopadhyay, R., Showkatian, E., Muneer, A., Zafar, A., Alvarez, F.R. et al. The next layer: augmenting foundation models with structure-preserving and attention-guided learning for local patches to global context awareness in computational pathology. *NPJ Precision Oncology* **10**, 13 (2026).

[18] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y. et al. Advances in medical image analysis with vision Transformers: A comprehensive review. *Medical Image Analysis* **91**, 103000 (2023).

[19] Zhou, C., Xu, J., Prakash, R., Torres-Cabala, C.P., Chen, C.B., Madduri, K. et al. Developing deep learning pipeline of Whole-Slide images for enhanced Diffuse Large B Cell lymphoma (DLBCL) subtyping and outcome prediction: Leveraging Self-Attention Transformer for training and inference. *Blood* **142** (Suppl 1), 904 (2023).

[20] Atabansi, C.C., Nie, J., Liu, H., Song, Q., Yan, L., Zhou, X. A survey of Transformer applications for histopathological image analysis: New developments and future directions. *BioMedical Engineering OnLine* **22**, 96 (2023).

[21] Campanella, G., Kwan, R., Fluder, E., Zeng, J., Stock, A., Veremis, B. et al. Computational Pathology at Health System Scale – Self-Supervised Foundation Models from Three Billion Images. *arXiv preprint* arXiv:2310.07033 (2023).

[22] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. *Lecture Notes in Computer Science* **13803**, 205–218 (2023).

[23] Ahmed, A., Zeng, X. A Robust Framework for Domain-Generalized Classification of Ovarian Cancer Histology images. *Diagnostics* **15**, 2954 (2025).

[24] Rahman, M.I. Fusion of vision transformer and convolutional neural network for explainable and efficient histopathological image classification in Cyber-Physical healthcare systems. *Journal of Transformative Technologies and Sustainable Development* **9**, 1 (2025).

[25] Kuo, R., Soltan, A.A.S., O'Hanlon, C., Hasanic, A., Clifton, D.A., Collins, G. et al. Benchmarking transformer-based models for medical record de-identification in a single center multi-specialty evaluation. *iScience* **28**, 113732 (2025).

[26] Zhao, J., Wen, X., Ma, L., Su, L. Automated classification of lymphoma subtypes from histopathological images using a U-Net Deep Learning Model: Comparative Evaluation study. *JMIR Medical Informatics* **14**, e72679 (2026).

[27] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D. et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *arXiv preprint* arXiv:2102.12122 (2021).

[28] Springenberg, M., Frommholz, A., Wenzel, M., Weicken, E., Ma, J., Strodthoff, N. From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. *Medical Image Analysis* **87**, 102809 (2023).

[29] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X. et al. A method for normalizing histology slides for quantitative analysis. *IEEE International Symposium on Biomedical Imaging (ISBI)* 1107–1110 (2009).

[30] Franco, D., Navarin, N., Donini, M., Anguita, D., Oneto, L. Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neurocomputing* **470**, 318–334 (2021).

[31] Mersha, M., Lam, K., Wood, J., AlShami, A.K., Kalita, J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* **599**, 128111 (2024).

[32] Kashefi, R., Barekatain, L., Sabokrou, M., Aghaeipoor, F. Explainability of vision transformers: a comprehensive review and new perspectives. *Multimedia Tools and Applications* **85**, 2 (2026).

[33] Rainio, O., Teuho, J., Klén, R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports* **14**, 6086 (2024).

[34] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S. et al. Transformers in medical imaging: A survey. *Medical Image Analysis* **88**, 102802 (2023).

[35] Zeng, A., Chen, M., Zhang, L., Xu, Q. Are transformers effective for time series forecasting? *arXiv preprint* arXiv:2205.13504 (2022).

[36] Chen, C.L., Chen, C.C., Yu, W.H., Chen, S.H., Chang, Y.C., Hsu, T.I. et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature Communications* **12**, 1193 (2021).

[37] Dehkharghanian, T., Bidgoli, A.A., Riasatian, A., Mazaheri, P., Campbell, C.J.V., Pantanowitz, L. et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic Pathology* **18**, 67 (2023).

[38] Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications* **12**, 4423 (2021).

[39] Zhang, Z., Zhang, H., Zhao, L., Chen, T., Arik, S.O., Pfister, T. Nested Hierarchical Transformer: towards Accurate, Data-Efficient and Interpretable Visual understanding. *arXiv preprint* arXiv:2105.12723 (2021).

[40] Tang, G.K., Lim, C.C., Hussain, F.A., Oung, Q.W., Yajid, A.I., Mohammad Azmi, S. A Novel Hybrid Pyramid ViT-ConvFormer Architecture with Dynamic Early Stopping Analysis for Enhanced Diagnostic Efficiency in Diffuse Large B-Cell Lymphoma. *Zenodo* (2026). https://doi.org/10.5281/zenodo.21866406.