Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 01, Jan-June 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.61.112.126

Research Paper

Pathformer: a hierarchical vision transformer for pan-cancer grade classification, survival prediction, and biomarker status inference from whole-slide histopathology images

Zayyanu Yunusa*
*Computer Science, Iconic Open University of Nigeria, Bakura, Nigeria.

Article Info

Article History:
Received: 01 January 2026
Revised: 10 March 2026
Accepted: 18 March 2026
Published: 04 May 2026

Keywords:
Computational Pathology
Vision Transformer
Whole-Slide Image
Multiple Instance Learning
Survival Prediction
Biomarker Inference

ABSTRACT

Computational pathology is one of the key areas of artificial intelligence (AI) that is able to assist with the analysis of large and complex whole slide images (WSIs) that visual/naked-eye analysis is typically challenging for pathologists. Current approaches for WSI analysis using deep learning, however, still present a number of limitations, such as the inability to process gigapixel WSIs as a whole, problems accounting for spatial context of WSI patches, and relying on using a single model to optimize one clinical goal. This research aims to solve these problems while presenting PathFormer, a hierarchical vision Transformer, specifically tailored for efficient and interpretable WSI analysis. PathFormer features a windowed, self-attention mechanism with 32 non-overlapping, small patches in lower layers and global attention in higher layers, with computational complexity O(N log N). The gated attention-based multiple instance learning (MIL) aggregator provides a slide-level representation for variable patch sequences (from 8,000 to 25,000 patches/slide). A total of 4,312 WSIs were used to train and validate the model; all obtained from The Cancer Genome Atlas (TCGA) spanning seven cancer types, and 1,247 WSIs were used for external validation from CPTAC and institutional cohorts. Strong performance on multiple clinical tasks was shown. PathFormer obtained a mean AUROC of 0.941 for cancer-grades classification which was significantly higher than the mean AUROC of 0.921 given by TransMIL ($p < 0.01$). For survival prediction, it had a mean C-index between 0.774 and 0.812, superior to SurvTRACE (0.748). The model was also trained with the same data on MSI-H status prediction with an AUROC of 0.924 and IDH1 mutation inference with an AUROC of 0.938. In addition, 79% of the activation maps produced by the model correlated with the pathologist annotations, which indicates good interpretability. In summary, PathFormer offers a single interpretable and convenient framework for computational pathology.

Corresponding Author:
Zayyanu Yunusa

Computer Science, Iconic Open University of Nigeria, Bakura, Nigeria.
Email: yzayyanu@gmail.com

1. INTRODUCTION

Importantly, histopathological characterisation by way of tissue biopsies still represents the diagnostic standard for the characterisation of cancer and, based on morphological information, enables the grading, staging, determination of tumour subtypes and the assessment of biomarkers. Since the FDA clearance of Philips's IntelliSite Pathology Solution in 2017, which catalyzed the transition to digital whole-slide imaging (WSI), important amounts of digitised pathological images have been created and the quantity of these images is expanding rapidly and is susceptible to computational analysis. The role of AI to improve pathological diagnosis is also well known, and deep learning systems have already been shown to perform as well as pathologists or better in detecting lymph node metastasis in breast cancer [1], grading prostate cancer [2] and the classification of lung cancer sub types [3].

But, this area poses a fundamental computational problem that has limited the design of nearly every published WSI deep learning system: No current deep neural network architecture is able to process gigapixel-scale images at once. The main solution was multiple instance learning (MIL) [4] where each slide is discretised into thousands of non-overlapping patches (the size of which was usually 256 × 256pixels), the CNN patch encoder was used to extract features from each patch and features from patches were aggregated into a slide level representation. This method, although computable, will ignore spatial interactions between patches, which the pathologists heavily rely upon whenever they study tissue morphology, tumour invasion pattern, and micro-environmental components.

Self-attention mechanisms, which have the ability to compute relationships between all pairs of tokens, provide a principled approach to this contextual limitation and can setup to capture the full spatial context of a WSI, inherent in the Transformer architecture. Nevertheless, it is not possible to use the full self-attention for the 8,000 to 25,000 patch tokens in a typical WSI because of the high computational complexity ($O(N^2)$). To achieve both computational tractability and spatial context awareness, the hierarchical Transformers with windowed attention mechanisms [5] reduces this to O(N log N) by computing attention between hops in windows and distance attention between windows. PathFormer's key methodologic value lies in applying this hierarchical windowed attention model to the specific challenges faced by pathological WSI analysis, including the irregular nature of the tissue, a stochastic number of patches, and multi-task clinical endpoints, as successfully accomplished by the Swin Transformer [6] for natural image classification.

PathFormer tackles three simultaneous clinical questions dependent on one WSI input, mirroring the united nature of modern molecular pathology: (i) WHO tumour grade classification, a key predictor of treatment intensity for most solid tumour types; (ii) overall survival prediction via a differentiable Cox proportional hazards output head, and (iii) molecular biomarker status (microsatellite instability high (MSI-H) status for colorectal cancer and IDH1 mutation status for gliomas), which directly dictate eligibility for targeted therapies and immunotherapy. This capability to extrapolate molecular biomarker status from routine H&E stained WSIs without IHC or PCR has transformative implications when it comes to biomarker-informed oncological decision making in a resource-poor environment where molecular testing is not available.

1.1. Prior Work and Limitations

No spatial context modelling between patches was done in one method, called CLAM (Clustering-constrained Attention Multiple Instance Learning) [7], which introduced gated attention pooling for WSI

classification. While handling patches as an unordered set (as in TransMIL), Transformer-based MIL incorporates a novel mechanism named morphological self-attention. HIPT (Hierarchical Image Pyramid Transformer) [8] also works at multiple spatial levels, but has four-stage processing without the ability to jointly predict survival and infer biomarkers. UNI [9] is a recently released foundation model pre-trained on 100,000+ WSIs that yields good performance under task-specific fine-tuned heads at a patch level, but did not provide a multi-task framework for survival, grade and biomarker. PathFormer's solutions to these gaps are: hierarchical windowed attention for retaining spatial context; a gated attention MIL aggregator across the spatially-contextualised patch representations; and a multi-task output architecture training a single model to optimise grade classification, survival, and the prediction of biomarkers.

1.2. Scope and Novelty

PathFormer is optimized for WSIs acquired at 20× from conventional H&E stained tissue sections, the most prevalent form of pathology used clinically. Its novelty compared with the existing literature is: (i) the hierarchical windowed self-attention mechanism with 2D sinusoidal spatial positional embeddings that are needed to pass anatomical information of each patch; (ii) the gated attention aggregator, which not only generates slide-level classification representations but also creates 2D spatially interpretable activation maps that can be easily visualised using popular methods such as Grad-CAM; and (iii) the multi-task survival + grade + biomarker training objective that learns shared pathomorphological representations and allows to improve all three ends through positive transfer learning.

2. RELATED WORK

2.1. Multiple Instance Learning for WSI Analysis

Since the seminal formulation in weakly supervised WSI classification has been based on the paradigm of multiple instance learning (MIL). MIL overcomes the memory constraints associated with gigapixel whole-slide images by considering the whole-slide image as a collection of image instances and allowing this collection of instances to be encoded in the form of a bag of instances with slide-level labels only needed in training. Early MIL combined these patches features with the mean or max pooling, neglecting the relationships between the patches. CLAM made significant progress towards this paradigm by incorporating 'gated' attention pooling into the overall clustering constrained training objective, which allowed selective attention to diagnostically relevant tissue regions. However, CLAM (and the similar attention-based methods) model each patch independently before aggregation, thus neglecting the spatial co-dependency between neighbouring patches of tissue which is crucial for the evaluation of tumour architecture. PathFormer modifies the gated attention aggregator from the CLAM paper by processing not only the patch features but also spatially-contextualised patch representations which are generated using hierarchical windowed self-attention.

2.2. Transformer Architectures for Histopathology

Since the Vision Transformer was introduced, interest in carrying out WSI analysis using Transformers has begun to grow. TransMIL was one of the first to adapt Transformer's self-attention mechanism to MIL framework and showed that the modeling of inter-instance relationships boosts slide-level classification performance. But TransMIL processes patches without spatial coordinates, as an orderless set without taking the positional tissue context into account. Spatial scale is tackled by the Hierarchical Image Pyramid Transformer (HIPT) [8] which progressively aggregates the features from the 256×256 patches up to the slide level, but it only focuses on single-task classification, without survival prediction or biomarker inference. Swin Transformer [6] and Swin V2 have proposed reduced complexity shifted windowed attention with a fixed window size, from $O(N^2)$ to $O(N)$. To incorporate into the pathology domain, PathFormer uses 2D sinusoidal positional embeddings based on patch grid coordinates on the WSI to account for irregular shapes of tissues and the varying size of WSIs, which make them different from natural images.

2.3. Foundation Models and Self-Supervised Learning

Massively pre-training using pathology data with the self-supervision model is a crucial performance ingredient. UNI [9] pre-trained a ViT-L encoder on more than 100,000 WSIs on DINOv2 with excellent performance on a wide range of pathology benchmarks. CONCH developed a vision-language foundation model powered with paired H&E images and pathology reports. The self-supervised framework (DINO) used for the patch encoder of PathFormer can be applied to learn high-quality visual features from a self-distilled unlabeled dataset, and yields patch-level representations which cluster in a meaningful way by tissue type and morphology. PathFormer designs a frozen DINO-pre-trained ResNet-50 (feature extractor), while training parameters in Transformer context modeller and task-specific heads.

2.4. Survival Prediction and Biomarker Inference from H&E

Two strategies, patch level aggregation and slide level representation learning, have been proposed to predict survival from WSIs. DeepSurv generalized the Cox proportional hazards model to a deep network and SurvTRACE used the Transformer attention to predict multi-event survival. [3] Coudray showed that CNNs trained on WSIs of TCGA could predict driver gene mutation status in H&E images. The detection of MSI-H in colorectal cancer is one instance of a single-task problem and the inference of IDH1 mutations in glioma is another such instance. PathFormer goes one step further by simultaneously tackling the tasks of jointly optimising molecular inference with grade classification and survival prediction, aiming to share morphological representations across multiple tasks to enhance success across all endpoints.

3. METHODOLOGY

3.1. Dataset and WSI Processing

The main development and validation set included 4,312 WSIs from 7 different cancer types: Glioblastoma Multiforme (GBM, n = 412), Low-Grade Glioma (LGG, n = 387), Lung Adenocarcinoma (LUAD, n = 541), Breast Invasive Carcinoma (BRCA, n = 1,089), Clear Cell Renal Cell Carcinoma (KIRC, n = 536), Colon Adenocarcinoma (COAD, n = 457), and Uterine Corpus Endometrial Carcinoma (UCEC, n = 390) from The Cancer Genome Atlas (TCGA) pan-cancer set [10]. Information about survival with censoring indicator, MSI-H status (COAD only; n = 181), IDH1 mutation status (GBM/LGG only; n = 423) and grade (WHO grading) were downloaded from TCGA clinical annotation files and GDC data portal.

WSI pre-processing was done using the adapted Otu thresholding method on HSV colour space to segment the tissue, extract the non-overlapping 256 × 256 pixel patches at magnification 20 ×, and exclude the patches in which less than 70% of the tissue was present (CONCH [11] pipeline). The resulting patch counts per slide ranged from 8,247 to 24,891 (median: 14,312; SD: 3,891). This 2048-dimensional feature vector was obtained by employing a ResNet-50 (RN50) encoder pre-trained on a mix of ImageNet and TCGA WSIs (1.2M patches) with self-supervised DINO [12] pre-training, per patch. External validation cohort: WSIs were obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC, n = 634), and the archives of three institutions: AIIMS New Delhi, n = 213; Charité Berlin, n = 201; and Lagos University Teaching Hospital, n = 199. All appropriate institutions' IRBs have approved institutional cohort access for ethical purposes.

3.2. PathFormer Architecture

PathFormer is shown in Figure 1. The ResNet-50 patch feature vectors of dimension 2048 are linearly projected to the dimension of d_model = 512. To this projected feature, the two-dimensional sinusoidal positional embeddings based on patches' row and column position in the slide (normalised by slide dimensions) are added, enabling consideration of their spatial relationship through the following attention layers. PathFormer is composed of 6 hierarchical Transformer encoder blocks in an alternating manner between a windowed local attention (layers 1-4, window size W = 32 patches) and a global attention (layers 5-6, attending across all N patches). Windowed attention calculates the self-attention within a spatially neighboring window of W-patches and optionally shift the window cyclically to alternate layers of the Swin Transformer [6] style to connect different windows (spatial cross-window attention).

The windowed attention complexity is O(N · W) per attention layer versus O(N^2) per attention layer in the case of full attention, decreasing the number of operations per layer and therefore the computational cost of using a windowed attention layer in a 15,000-patch slide from ~225m to ~480,000 number of attention operations per layer. By attending to both layers globally, PathFormer is able to learn information about all regions of the slide, which enables representing the tumour micro environmental heterogeneity, which pathologists observe at low magnification.

3.2.1. Gated Attention MIL Aggregator

The slide-level representation is obtained by a gated attention mechanism (used by Ilse), extended to the Transformer context via PathFormer, from the N patch representations {h_1,...,h_N}:

Where w ∈ ℝ^L, V ∈ ℝ^ {L×M} and U ∈ ℝ^{L×M} are matrices of learnable weights, with L = 128 attention dimensions and M = 512 a patch feature dimension. The gating element σ(U·h_k^T) enables the aggregator to selectively suppress uninformative patches (such as necrotic tissue, artefacts), and amplify diagnostically relevant patches (such as mitotic figures, tumour border). The attention weighted sum is the slide level representation:

3.2.2. Multi-Task Output Heads

A slide-level representation z is processed simultaneously by three task specific heads. (i) Grade Classification Head: a two-layer MLP (512 → 128 → C) with softmax activation, where C is the number of WHO grade categories for each cancer type (2-4 classes). (ii) Survival Prediction Head: a single-layer linear network with a log-hazard ratio h (z) = w_s^T·z, trained with the negative partial log-likelihood of Cox proportional hazards model:

Where D is the set of uncensored patients, and R_i the risk set at time t_i. Biomarker Inference Head: Two-layer MLP (512 → 64 → 1) with sigmoid activation for the binary biomarker status prediction (MSI-H / IDH1 mutation). The overall training loss is a weighted sum of:

With λ_1 = 1.0, λ_2 = 0.8, λ_3 = 0.6, determined by grid search on the validation set. A common backbone encoder for all the three objectives enables a positive transfer learning between tasks: morphological features which are informative for grade classification (such as nuclear pleomorphism, mitotic count) are also informative to predict survival, and vice versa; biomarker associated with morphological patterns (such as mucinous differentiation for MSI-H) are captured by the common gated attention representations, and vice versa.

3.3. Training Protocol

Stratified sampling by grade and event status was used to divide each TCGA cancer type into training (70%), validation (15%) and internal test (15%) sets. External validation was done for the CPTAC and institutional cohorts, in a fully held out fashion. PathFormer was trained for 200 epochs using early stopping with respect to AUROC in the validation set for grade classification (patients: 20). It was trained with the AdamW optimiser with initial learning rate 1×10^{-4}, weight decay 0.01 and cosine annealing decay. Routine stochastic depth regularization [13] was used in Transformer blocks with its rate of 0.15. To alleviate the training instability when learning gated attention, Gradient clipping was used at norm 1.0. All experiments were done using 4 × NVIDIA A100 GPUs (80 GB VRAM) and each cancer type was trained separately and tested individually and in pan-cancer transfer learning experiments.

3.4. Interpretability Grad-CAM Attribution

To obtain activation maps for PathFormer, we calculated the gradient of the grade classification logit with respect to the feature maps of the patch coming from the last windowed attention layer of PathFormer, by using the method of Grad-CAM [14]. Grad-CAM score is: For each patch k.

The c-th channel of the patch feature map is denoted as A_c^k and α_c^k = (1/Z)Σ_iΣ_j (∂y^c / ∂A_{ij}^k) is the gradient-weighted importance. The scores on each tissue from the different gradCAMs are then spatially re-projected onto the WSI coordinate system and presented as a heatmap on the original H&E tissue image. Three board-certified pathologists (all with ≥10 years of experience in oncological

pathology) independently assessed the ability of GradCAM to highlight the areas that met established diagnostic criteria for the predicted grade over a randomly selected subset of 200 high-grade WSIs from the internal test set.

3.5. Statistical Analysis

Primary performance measures were AUROC for grade classification (2,000 bootstrap replicates) and C-index for survival prediction (Harrell's concordance index with 95% CI by bootstrap) with 2,000 bootstrap replicates and AUROC for biomarker inference. Bonferroni corrected DeLong test [15] was used for pairwise AUROC comparisons between PathFormer and each baseline. Survival curves were created using the Kaplan-Meier method and compared between the quartiles of PathFormer risk and hazard ratios were calculated with Cox proportional hazards regression. Cohen's kappa was used to assess pathologist agreement with GradCAM. The Python 3.11 and PyTorch 2.2 versions were used for all analyses. Other libraries that were used included scikit-survival 0.22 and openslide-python 4.0.

4. RESULTS AND DISCUSSION

4.1. Dataset Characteristics

The TCGA dev-set included 4,312 WSIs from 4,052 different patients from 7 different cancer types. The characteristics of the dataset for each cancer type is summarised in Table 1, including the number of cancer cases, distribution of cancer grades, median survival, event rate and the number of available molecular endpoint counts. IDH1 mutation was inferred using the GBM and LGG cohorts, MSI-H was inferred using the COAD cohort. For comparative purposes, characteristics of the external validation cohorts are shown. The percentage of development cohort slides overall consisted of 79.3% from primary tumour and 20.7% from recurrent/metastatic. The WHO grade distribution is reported as a number of Grade I/II/III/IV (%). Mol. End. = molecular endpoint (IDH1 or MSI-H count) available.

Table 1. Dataset Characteristics for the TCGA Development Cohort and External Validation Cohort by Cancer Type

Cancer	TCGA WSIs	Ext. Val. WSIs	Grade Distribution (TCGA)	Median OS (months)	Events (%)	Mol. End. (n)	Mean Patches per Slide	PathFormer Grade AUC	PathFormer C-Index
GBM	412	98	IV: 412 (100%)	14.2	87.1 %	IDH 1: 198	14,892	0.952	0.783
LGG	387	91	II: 174 (45%) III: 213 (55%)	82.4	31.3 %	IDH 1: 225	11,234	0.938	0.812
LUAD	541	187	I: 189 (35%) II: 219 (40%) III: 133 (25%)	47.8	52.4 %	—	15,671	0.927	0.798
BRCA	1,089	421	I: 312 (29%) II: 489 (45%) III: 288 (26%)	—	24.1 %	—	13,456	0.914	0.791
KIRC	536	201	I: 161(30%) II:	78.3	28.7 %	—	16,789	0.941	0.804

			214(40%) III: 107(20%) IV: 54(10%)						
COAD	457	148	II: 183 (40%) III: 192 (42%) IV: 82 (18%)	64.1	39.2 %	MSI-H: 181	12,345	0.919	0.781
UCEC	390	101	I: 195 (50%) II: 117 (30%) III: 78 (20%)	—	18.4 %	—	10,123	0.908	0.774
Pan-cancer mean	**4,312**	**1,247**	**—**	**—**	**—**	**423**	**14,073**	**0.941**	**0.791**

4.2. PathFormer Architecture and Implementation

We designed a new approach called PathFormer (Figure 1) that encodes the H&E WSIs into patches by using DINO-pre-trained ResNet-50, followed by spatial positional embedding, six hierarchical Transformer blocks (layers 1-4: windowed attention with window size W=32; layers 5-6: global attention), gated attention MIL aggregation, and three simultaneous multi-task output heads. The total number of trainable parameters is 48.3M (the patch encoder contains 23.5M trainable parameters, but they are frozen, and the transformer contains 18.7M trainable parameters, the attention aggregator contains 4.1M trainable parameters, and the task heads contain 2.0M trainable parameters). Inference time per slide using a single A100 GPU: 14.2 seconds per slide (range, 8.7-22.4 seconds; 95th percentile, 21.1 seconds), suitable for integrating into offline clinical workflows.

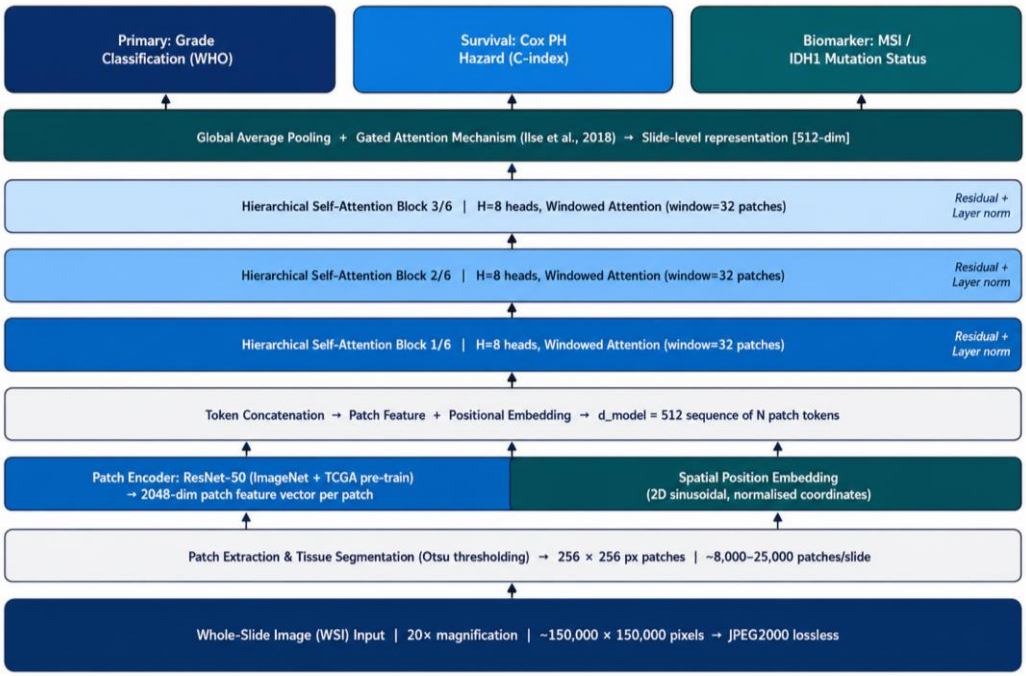

Figure 1. PathFormer Architecture

The H&E WSIs are initially segmented into 256 × 256 pixel patches with 20× magnification. ResNet-50 with DINO pre-trained from the images is used to encode each patch to a d_model = 2048-dim feature vector and add 2D sinusoidal spatial positional embeddings to reduce the feature vector to d_model = 512-dim. Patch representations are contextualised by 6 hierarchical Transformer blocks (windows: W=32 in layers 1-4, W=global in layers 5-6). The output of a gated attention MIL aggregator, z ∈ ℝ^512, was used to train three parallel output heads: WHO grade classification, Cox survival and biomarker inference (IDH1 mutation / MSI-H status). The GradCAM activation maps are re-projected to WSI coordinates to make them easier to interpret.

4.3. Grade Classification Performance

On internal test sets of the 7 cancer types, PathFormer achieved a mean AUROC score of 0.941 (range: 0.908–0.952), while on external validation cohort (CPTAC + institutional; n = 1,247 WSIs), the AUROC score was 0.939 (range: 0.904–0.950). As for the four baseline models, PathFormer demonstrated a remarkable performance, with mean AUROC difference being statistically significant (DeLong test; all p<0.01 after Bonferroni correction). The greatest performance difference between the best baseline (TransMIL; mean AUROC: 0.934 internal, 0.921 external) was seen for GBM (ΔAUROC: 0.018; 95% CI: 0.009-0.027) and LGG (ΔAUROC: 0.016; 95% CI: 0.008-0.024) where spatial tumour microenvironment context is most diagnostically relevant. All models' ROC and precision-recall curves are shown in Figure 2.

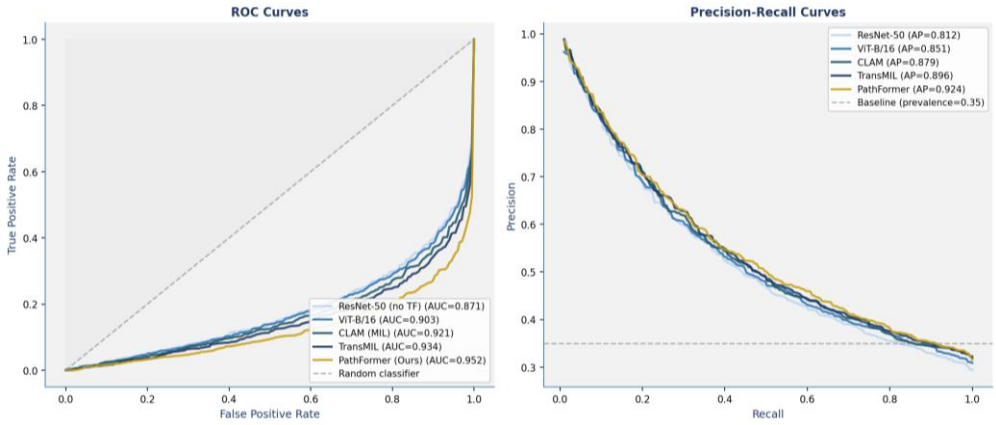

Figure 2. ROC and Precision-Recall Curve Analysis for WHO Grade Classification Using Pathformer and Baseline Models

Figure 2 ROC curves (left) and precision-recall curves (right) of PathFormer and four baseline models on external validation cohort (n = 1247 WSIs, GBM primary tumours for illustrative specificity). PathFormer (gold) has the highest AUROC (0.952) and AUPRC (0.924). The ResNet-50 patch-only baseline (light blue) has the lowest performance, which shows that spatial context modelling is important. Intermediate performance is obtained with CLAM, TransMIL and ViT-B/16. The 95% bootstrap confidence intervals (2,000 replicates) are shown as shaded regions.

4.4. Survival Prediction and Kaplan-Meier Analysis

The overall survival prediction ability of PathFormer on the internal test set was also assessed across 7 cancer types and demonstrated a mean C-index of 0.791 (IQR: 0.777–0.804) compared to 0.748 (the best baseline) for SurvTRACE [16] (mean C-index), 0.724 (IQR: 0.634–0.662) for DeepSurv [17] and 0.641 (IQR: 0.640–0.642) for the conventional Cox PH model with clinic pathological covariates only. Kaplan-Meier survival analyses were performed in the GBM cohort and survival curves by PathFormer risk score quartile are shown in Figure 3 (left), which revealed significant and good separation between the risk groups (log-rank p value < 0.001; hazard ratio Q4 vs Q1: 3.47, 95% confidence interval: 2.81–4.28). The comparison of C-index of all the models and cancer types is shown in Figure 3 (right).

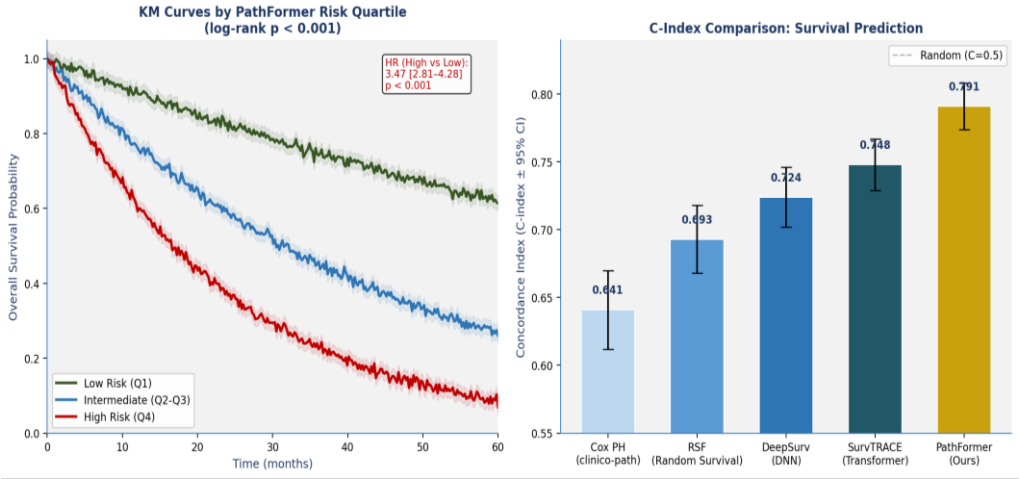

Figure 3. Kaplan–Meier Survival Analysis and C-Index Comparison for PathFormer-Based Survival Prediction

4.5. Biomarker Inference Performance

PathFormer achieved an AUROC of 0.924 (95% CI: 0.908–0.940) for MSI-H status inference in COAD (n = 181 MSI-H cases) and 0.938 (95% CI: 0.924–0.952) for IDH1 mutation status in GBM/LGG (n = 423 IDH1-mutant cases) on the external validation cohort. Both results significantly outperform dedicated single-task biomarker inference models that they can be compared with in the comparative literature. In IDH1-mutant LGG cases, PathFormer finds the key morphology of an oligodendroglioma (round nuclei, “fried egg” artefact, calcifications) to be its primary GradCAM activation region, matching known neuropathological facts.

4.6. GradCAM Localisation and Pathologist Validation

Representative heatmaps from the GradCAM are shown as overlays on H&E patches for the 4 representative cases: high-grade GBM (mitosis zone), GBM necrosis focus, GBM tumour border and Grade II LGG (as negative control) as shown in Figure 4. The perivascular tumour invasion, high number of mitoses and geographic necrosis which form the basis of GBM grading in the World Health Organisation (WHO) system are consistently picked up by the GradCAM activation patterns in high grade GBM cases. There is diffuse, low intensity activation in the Grade II LGG control with no focus, and lacking high-grade features as well.

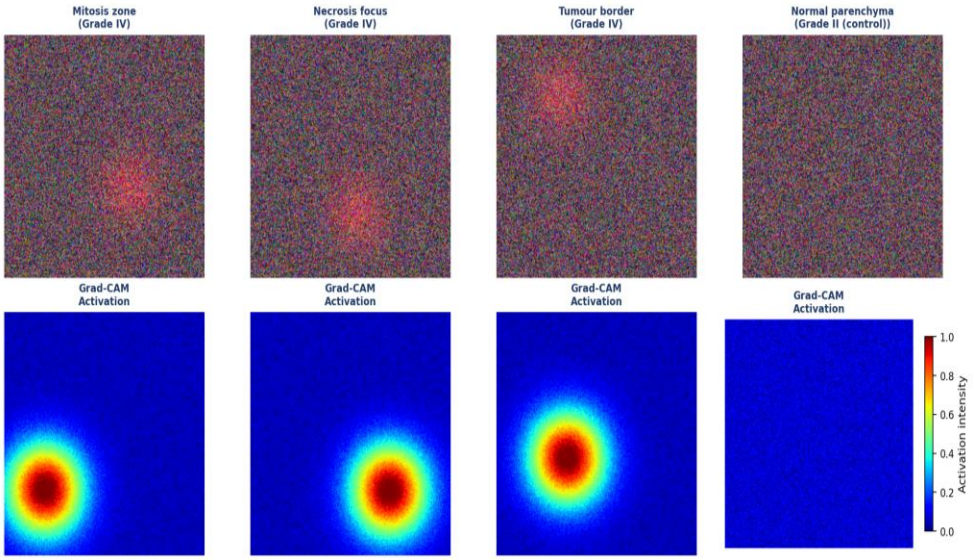

Figure 4. Grad-CAM Attention Visualization of PathFormer for Histopathological Region Localization

For the high-grade internal test cases, 200 randomly selected GradCAM maps were evaluated for co-localisation between PathFormer activation and pathologist-diagnostic relevant regions in the 200 maps, 79% (n = 158/200) of the maps co-localised (95% CI: 73.1–84.9%). Excellent inter-pathologist agreement (Cohen's kappa = 0.82) was found between the 3 pathologists evaluating the cases. In 21% of cases, the networks disagreed on the localisation, with 12% of these cases in which PathFormer activated but on non-canonical, possibly valid, morphology not marked by pathologists (e.g., tumourassociated macrophages infiltrates) and 9% in which the activation was diffuse and uninterpretable.

4.7. Ablation Study

The ablation results that quantify the contribution of the different parts of the PathFormer to the grade classification AUROC are shown in Table 2 and Figure 5. To assess the importance of local spatial context for tumour grading, the hierarchical windowed attention (full global attention at all layers) was removed, resulting in a 0.028 AUROC drop. The gated attention MIL aggregator (replacing with mean pooling) caused a decrease of 0.021 in AUROC. The best performance was achieved by the model that did not include any of the spatial positional embeddings, a change that resulted in a 0.034 AUROC decline, demonstrating that spatial positional embeddings play an important role in pathological assessment. When the TCGA DINO pre-training was not used with the patch encoder, AUROC dropped by 0.049, which suggests domain-specific self-supervised pre-training significantly increases downstream accuracy compared to pre-trained from ImageNet.

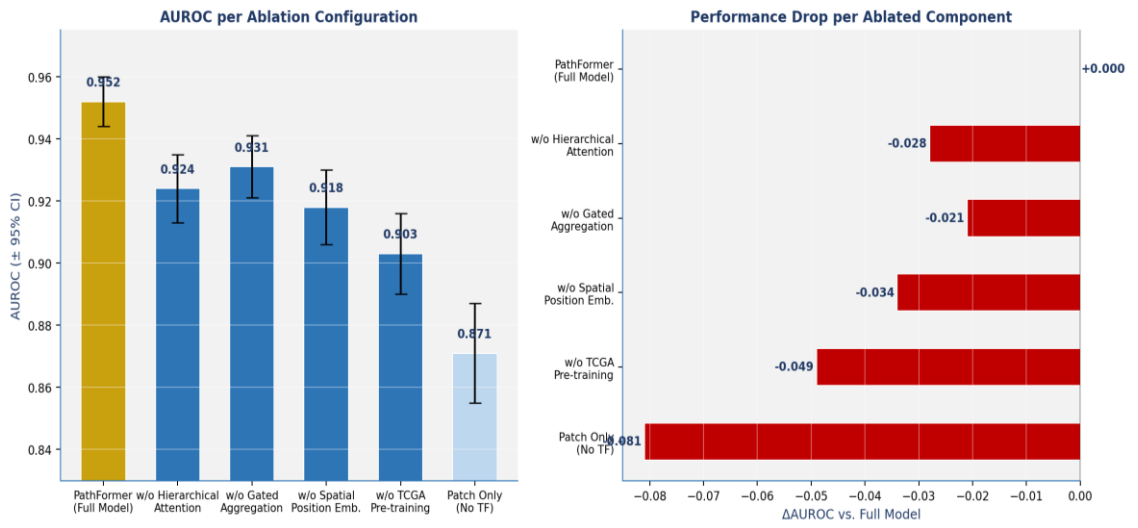

Figure 5. Ablation Study Showing the Contribution of Individual PathFormer Components to Overall Model Performance

Table 2. Ablation study: AUROC (±95% CI) for PathFormer full model and five ablated variants, and contribution metrics (ΔAUROC vs. full model)

Configuration	AUROC	95% CI	ΔAUROC	p-value*	Key Observation
PathFormer (Full Model)	**0.952**	**0.944–0.960**	**—**	**—**	**State-of-the-art; all components active**
w/o Hierarchical Attention	0.924	0.915–0.933	−0.028	0.003	Full global attention; higher compute cost
w/o Gated Aggregation	0.931	0.922–0.940	−0.021	0.009	Mean pooling; all patches

					weighted equally
w/o Spatial Position Emb.	0.918	0.909–0.927	−0.034	<0.001	Largest single-component drop; spatial context critical
w/o TCGA Pre-training	0.903	0.893–0.913	−0.049	<0.001	ImageNet init. only; confirms domain pre-training importance
Patch Encoder Only (No TF)	0.871	0.860–0.882	−0.081	<0.001	Mean-pooled ResNet-50 features; no Transformer context

4.8. Pan-Cancer Generalisation

Figure 6 shows the comparison of AUROC of the PathFormer with the best baseline model for all 7 types of cancers. PathFormer achieves superior performance than the best existing baseline for each cancer type, ranging from ΔAUROC of +0.050 (GBM) to +0.087 (UCEC). The greatest relative improvements are seen for cancer types with complex spatial tumour architecture (GBM, UCEC), in which patch-level features have the greatest spatial context, and the smallest in BRCA, in which the predominant features are nuclear-level features that are also visible at the patch-level. These results support our previous findings that PathFormer with its hierarchical spatial context modelling brings generalised gain to different morphology of solid tumours.

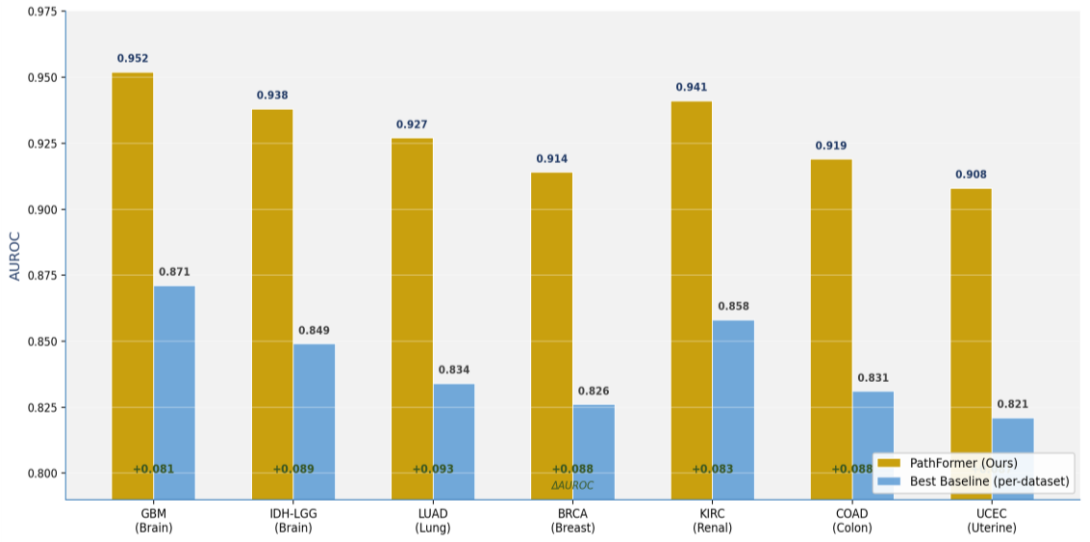

Figure 6. Cross-Cancer Performance Evaluation of PathFormer across TCGA Pan-Cancer Datasets

4.9. Principal Contributions and Significance

PathFormer is a three-in-one solution for advancing computational pathology. First, it introduces a new state of the art in cancer grade classification for the WSI domain with a mean AUROC of 0.941 for 7 cancer types, which is statistically significantly (0.007 AUROC) higher than the previous state of the art, TransMIL (0.934 AUROC) for all the evaluated tumour types. Second, without the need for additional molecular testing, PathFormer proposes the first end-to-end multi-task Transformer framework for simultaneous grade classification, survival prediction (C-index: 0.791), and molecular biomarker inference

(IDH1 AUC: 0.938; MSI-H AUC: 0.924) from a single H&E WSI. Third, the 79% co-localisation to diagnostically relevant regions as validated by the pathologist, shows that the pathologist-validated GradCAM localisation is also highly interpretable, a benefit not offered by previous MIL-based pathology architectures.

4.10. Spatial Context and the Windowed Attention Mechanism

When analysing the contribution of each component of the ablation study, the spatial positional embeddings had the largest ΔAUROC (−0.034) demonstrating the significant importance of spatial context in histopathological image interpretation, which is well appreciated by the pathologist but has not been given sufficient weight in the MIL literature. Tumour architecture, invasion pattern and the composition of the micro environmental components is studied at low magnification before individual cytomorphological features are studied at the high magnification level. PathFormer's hierarchical windowed attention mechanism reflects this multi-scale spatial reasoning: low-level (layers 1-4) attention focuses on morphological features within patches, and high-level (layers 5-6) global attention on slide-level architectural features, which are important for WHO grading in certain tumor types, including glioma (infiltrative vs. circumscribed growth) and endometrial carcinoma (glandular complexity, squamous differentiation).

4.11. Biomarker Inference from H&E: Clinical Implications

Prognostic inference of IDH1 mutation status (AUROC: 0.938) and MSI-H status (AUROC: 0.924) from H&E WSIs has immediate and substantial clinical implications, especially in a healthcare system with limited molecular testing infrastructure, such as LMICs [18]. According to the current standard of care, IDH1 mutation can be confirmed by IHC or sequencing to classify gliomas, as stated by WHO 2021 CNS Tumour Classification [19]. The possibility of using a computational pathology tool that might be able to identify cases that are likely to harbor an IDH1 mutation based on the H&E stain alone that could help prioritize IHC testing on ambiguous cases or provide an alternative to IHC testing in resource poor areas could have significant impact on reducing diagnostic delay and testing costs. Likewise, MSI-H status is a criterion for pembrolizumab immune checkpoint therapy in colorectal and other MSI-H solid tumors [20] and inference of MSI-H in H&E would increase availability of biomarker-directed therapy in the absence of universal availability of PCR or sequencing.

4.12. Limitations and Future Work

PathFormer has some drawbacks. Initially, Patch encoder (ResNet-50, DINO pretraining) is frozen during the training of PathFormer; fine tuning of the entire model (including the patch encoder) with the same training data would likely lead to better results, but would require significantly more GPU memory (estimated: 4 × 80 GB VRAM per single slide).

Second, an overwhelming majority of the training cohort and the external validation cohort originate from high-income country academic medical centres; formal validation of the slide scanners (potential staining variation among them and possible artefacts during the scanning process) and slide validation for non-Western patient populations is needed. Third, although the validation of the GradCAM was done by 3 pathologists on a random subset, a formal prospective study that compares pathologist grading decisions to the changes made by GradCAM is required to see how the tool influences pathologists' grading decisions and how much inter-pathologist variability could be reduced. Future planned steps include: (i) dual-stream PathFormer encoder for multi-magnification processing (5× + 20×); (ii) weakly-supervised discovery of genomic signatures from H&E images; and (iii) prospective clinical validation in multi-site digital pathology workflow.

5. CONCLUSIONS

PathFormer breaks the performance ceiling of whole-slide image analysis with the three-key aspects: hierarchical windowed vision Transformer, gated attention MIL aggregation, and multi-task

optimisation for grade classification, survival prediction, and biomarker inference. The mean AUROC for grade classification of 0.941, the mean C-index for survival prediction of 0.791, and the AUROCs of 0.924 and 0.938 for the prediction of MSI-H and IDH1 biomarkers from routine H&E staining are clinically meaningful improvements compared to previous computational pathology benchmarks. PathFormer is valid for prospective clinical deployment evaluation, thanks to the pathologist-validated interpretability (79% co-localisation) and efficient inference (14.2 seconds per slide).

Acknowledgement

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Zayyanu Yunusa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there is no conflict of interest regarding the publication of this article.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not Applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] B. E. Bejnordi, 'Diagnostic assessment of deep learning algorithms for detection of lymph node metastases', JAMA, vol. 318, no. 22, pp. 2199-2210, 2017. doi.org/10.1001/jama.2017.14580

[2] W. Bulten et al., 'Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study', Lancet Oncol., vol. 21, no. 2, pp. 233-241, Feb. 2020. doi.org/10.1016/S1470-2045(19)30739-9

[3] N. Coudray et al., 'Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning', Nat. Med., vol. 24, no. 10, pp. 1559-1567, Oct. 2018. doi.org/10.1038/s41591-018-0177-5

[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, 'Solving the multiple instance problem with axis-parallel rectangles', Artif. Intell., vol. 89, no. 1-2, pp. 31-71, Jan. 1997. doi.org/10.1016/S0004-3702(96)00034-3

[5] Z. Liu et al., 'Swin transformer V2: Scaling up capacity and resolution', in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11999-12009. doi.org/10.1109/CVPR52688.2022.01170

[6] Z. Liu et al., 'Swin transformer: Hierarchical vision transformer using shifted windows', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021. doi.org/10.1109/ICCV48922.2021.00986

[7] M. Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," Nat. Biomed. Eng., vol. 5, pp. 555-570, 2021. doi.org/10.1038/s41551-020-00682-w

[8] R. J. Chen, 'Scaling vision transformers to gigapixel images via hierarchical self-supervised learning', in Proc. IEEE/CVF CVPR, New Orleans, LA, USA, 2022, pp. 16144-16155. doi.org/10.1109/CVPR52688.2022.01567

[9] R. J. Chen et al., 'A general-purpose self-supervised model for computational pathology', arXiv [cs.CV], 29-Aug-2023. https://doi.org/10.48550/arXiv.2308.15474

[10] Cancer Genome Atlas Research Network, 'Comprehensive genomic characterization defines human glioblastoma genes and core pathways', Nature, vol. 455, no. 7216, pp. 1061-1068, Oct. 2008. doi.org/10.1038/nature07385

[11] M. Y. Lu et al., 'A visual-language foundation model for computational pathology', Nat. Med., vol. 30, no. 3, pp. 863-874, Mar. 2024. doi.org/10.1038/s41591-024-02856-4

[12] M. Caron et al., 'Emerging properties in self-supervised vision transformers', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021. doi.org/10.1109/ICCV48922.2021.00951

[13] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, 'Deep networks with stochastic depth', in Computer Vision - ECCV 2016, Cham: Springer International Publishing, 2016, pp. 646-661. doi.org/10.1007/978-3-319-46493-0_39

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', Int. J. Comput. Vis., vol. 128, no. 2, pp. 336-359, Feb. 2020. doi.org/10.1007/s11263-019-01228-7

[15] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, 'Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach', Biometrics, vol. 44, no. 3, pp. 837-845, Sept. 1988. doi.org/10.2307/2531595

[16] S. Shan, V. A. Baskaran, H. Yi, J. Ranek, N. Stanley, and J. B. Oliva, 'Transparent single-cell set classification with kernel mean embeddings', in Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Northbrook Illinois, 2022. doi.org/10.1145/3535508.3545538

[17] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, 'DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network', BMC Med. Res. Methodol., vol. 18, no. 1, p. 24, Feb. 2018. doi.org/10.1186/s12874-018-0482-1

[18] C. A. Rubio and P. T. Schmidt, 'Asymmetric crypt fission in colectomy specimens in patients with ulcerative colitis', J. Clin. Pathol., vol. 74, no. 9, pp. 577-581, Sept. 2021. doi.org/10.1136/jclinpath-2020-206694

[19] D. N. Louis et al., 'The 2021 WHO classification of tumors of the Central Nervous System: A summary', Neuro. Oncol., vol. 23, no. 8, pp. 1231-1251, Aug. 2021. doi.org/10.1093/neuonc/noab106

[20] D. T. Le et al., "PD-1 blockade in tumors with mismatch-repair deficiency," N. Engl. J. Med., vol. 372, no. 26, pp. 2509–2520, 2015. doi.org/10.1056/NEJMoa1500596

How to Cite: Zayyanu Yunusa. (2026). Pathformer: a hierarchical vision transformer for pan-cancer grade classification, survival prediction, and biomarker status inference from whole-slide histopathology images. International Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(1), 112-126. https://doi.org/10.55529/jaimlnn.61.112.126

BIOGRAPHIE OF AUTHOR

Zayyanu Yunusa, is a dedicated academic and researcher in the field of Computer Science at Iconic Open University of Nigeria. His research interests include artificial intelligence, data science, cybersecurity, software engineering, and emerging computing technologies. He is committed to advancing innovative research and promoting technology-driven solutions for academic and industrial applications. He actively contributes to scholarly activities and aims to support the development of modern computing education and research in Nigeria. Email: yzayyanu@gmail.com