Mitigating Shortcut Learning in Brain Tumour MRI Classification

Riya Basak
Department of Computer Science
University of Hertfordshire

Code: github.com/AnnyaB/HybridResNet50V2-RViT
Website: annyab.github.io/pfd-gste-research-article/

Abstract

Shortcut learning in medical imaging occurs when a classifier exploits acquisition, background or annotation cues that correlate with the label but do not define the pathology. This paper studies that failure mode in four-class brain tumour magnetic resonance imaging (MRI) classification. The model couples a 50-layer version-2 residual network (ResNet50V2) feature extractor from the PyTorch Image Models (`timm`) library with a Rotation-Invariant Vision Transformer (RViT). It introduces Pathology-Focused Disentanglement (PFD), which learns a soft spatial gate without segmentation masks, and Guided Semantic Token Evolution (GSTE), which transfers the gate to transformer tokens. Hybrid A guides 49 convolutional neural network (CNN) feature tokens while retaining an ungated CNN descriptor. Hybrid B guides both the CNN descriptor and up to 196 raw-image patch tokens, with bounded grid reduction when the gate is spatially concentrated. The corresponding ablations remove PFD and GSTE from each family. A Secure Hash Algorithm 1 (SHA-1) audit removes 297 exact duplicates from 7,023 files, leaving fixed training, validation and held-out test sets of 4,353, 1,089 and 1,284 images. Test accuracy ranges from 98.52% to 99.22%. Ablation B gives the highest classification result, with 99.22% accuracy and a 0.9920 macro-averaged F1 score (macro-F1). Hybrid B reaches 98.52% accuracy and 0.9849 macro-F1, but is the only model with tumour-centred Gradient-weighted Class Activation Mapping++ (Grad-CAM++) in all three reported cases. All four models remain highly confident on one shared subtype error. The experiment exposes a discrepancy between classification and qualitative spatial evidence: PFD–GSTE alters the observed evidence pathway without improving the best held-out classification result. These results evaluate a mitigation-oriented intervention rather than establish causal removal of shortcut learning; neither classification performance, qualitative saliency nor Monte Carlo (MC) dropout confidence alone is sufficient evidence of shortcut elimination.

1 Introduction

Brain tumours comprise heterogeneous diseases whose diagnosis and management depend on tumour type, location, molecular characteristics and imaging context [1–3]. Magnetic resonance imaging (MRI) provides detailed soft-tissue contrast, but tumour appearance varies with sequence, orientation, slice position and acquisition setting [4, 5]. Such variation complicates classification and creates correlations that need not arise from tumour tissue.

Shortcut learning occurs when a model obtains the correct output from an unintended predictive cue [6]. In medical imaging, such cues may arise from scanner characteristics, hospital-specific processing, visible markers, borders, background structure or dataset construction [7–10]. Accuracy on a familiar split does not identify the image evidence supporting a prediction. Predictive correctness and spatial evidence must therefore be evaluated separately.

Convolutional neural networks (CNNs) supply a local inductive bias and support transfer learning on moderate-sized medical datasets [11–13]. Vision Transformer (ViT) architectures model longer-range relations between image regions, but require control of data, regularisation and token cost [14, 15]. Krishnan et al. introduced the Rotation-Invariant Vision Transformer (RViT) for brain tumour MRI by combining rotated patch embeddings before transformer encoding [16]. The present model combines this rotation-invariant construction with a `timm` ResNet50V2 feature extractor and tests pathology guidance at two intervention sites.

This paper introduces Pathology-Focused Disentanglement (PFD) and Guided Semantic Token Evolution (GSTE) within two ResNet50V2–RViT families. PFD learns a one-channel gate from the deepest ResNet50V2 feature map, and GSTE converts that gate into mean-normalised token weights. Hybrid A confines guidance to 49 CNN-derived transformer tokens. Hybrid B applies the same gate to the CNN descriptor and a 196-patch image-token pathway, with optional grid reduction. Removing PFD and GSTE from each family gives the corresponding ablations.

The research question is: *to what extent can a pathology-guided CNN–Transformer intervention preserve held-out four-class classification performance while shifting the observed spatial evidence towards tumour-related regions, and what residual failure modes remain visible under family-specific controls?*

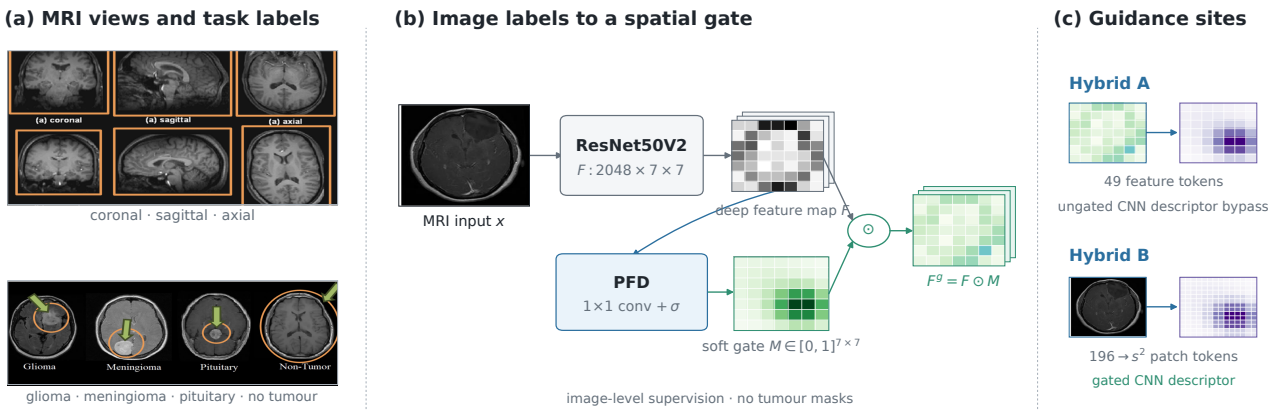


Figure 1: **Task formulation and pathology-guided intervention.** (a) Representative coronal, sagittal and axial MRI views illustrate variation in orientation, position and field of view; the lower row shows the four classification labels: glioma, meningioma, pituitary and no tumour. The orientation panel is reproduced and adapted from Figure 7 of Gholipour et al. [5]; the class panel is reproduced and adapted from Figure 1 of Yazdan et al. [17]. (b) ResNet50V2 maps an input slice x to the deepest feature map F; PFD predicts the spatial gate M and forms $F^g = F \odot M$ using image-level supervision. (c) GSTE-A transfers M to 49 CNN feature tokens while the Hybrid A CNN descriptor bypasses the gate. GSTE-B transfers the upsampled gate to at most 196 image-patch tokens and also gates the Hybrid B CNN descriptor.

Contributions. The paper contributes:

- a reproducible ResNet50V2–RViT family that combines local CNN features with four-rotation transformer processing for four-class brain tumour MRI classification;
- PFD, a learnable spatial gate trained from image-level labels and requiring no tumour masks;
- GSTE-A for static guidance of 49 feature tokens, and GSTE-B for guidance and bounded reduction of a 196-token image grid;
- a leakage-aware benchmark created by exact-duplicate removal, fixed split construction and held-out checkpoint evaluation; and
- a four-model study using classification, error analysis, Gradient-weighted Class Activation Mapping++ (Grad-CAM++), Attention Rollout and Monte Carlo (MC) dropout.

The suffixes A and B identify the implementation used by Hybrid A and Hybrid B, respectively. Thus, PFD-A and GSTE-A denote the feature-token guidance path, whereas PFD-B and GSTE-B denote the joint CNN-descriptor and image-patch guidance path.

2 Related Work

2.1 Clinical setting, benchmark scope and shortcut learning

The four labels considered here–glioma, meningioma, pituitary and no tumour–compress substantial clinical heterogeneity. Gliomas span molecularly and histologically distinct entities; meningiomas vary with age and anatomical site; and pituitary tumours require endocrine as well as anatomical assessment [1–3, 18, 19]. Paediatric tumours and tumour-like lesions introduce further distinctions that are absent from the four-folder classification setting [20–22]. A single red–green–blue (RGB) converted two-dimensional MRI slice therefore cannot reproduce a clinical diagnostic pathway. It instead defines a controlled setting in which representation and classification behaviour can be examined.

Clinically richer resources expose what this abstraction omits. The Brain Tumor Segmentation (BraTS) benchmark provides multimodal three-dimensional MRI with expert tumour annotations, while The Cancer Imaging Archive contains heterogeneous patient-level imaging collections [23, 24]. The present experiment uses the Nickparvar Kaggle aggregate because its four class folders define the intended classification task within the available computational budget [25]. The experimental snapshot, however, contains no patient identifiers, sequence metadata or centre-level grouping. Exact Secure Hash Algorithm 1 (SHA-1) filtering can eliminate byte-identical train–test overlap, but cannot establish patient independence or detect adjacent slices and transformed near-duplicates. The resulting benchmark is therefore suitable for a controlled internal experiment, not for clinical validation.

This distinction is consequential because predictive accuracy does not identify the evidence on which a medical-image classifier relies. Shortcut learning arises when a model exploits a predictive variable that is correlated with the target in the training distribution but does not constitute the intended pathology [6, 9, 10, 26]. Recent work has made this failure mode increasingly explicit. Stanley et al. show with controlled brain-MRI experiments that disease-related and bias-related information may both be encoded within

convolutional representations, while also demonstrating that representation of a bias does not necessarily imply its use as a classification shortcut [27]. Tinauer et al. provide a complementary empirical result: skull-stripping can introduce brain-contour information that remains predictive after substantial image-content changes, producing a measurable shortcut in MRI classification [28]. At the dataset level, Drenkow et al. introduce Generalized Attribute Utility and Detectability-Induced bias Testing (G-AUDIT) to quantify potential shortcut relationships among task labels, sensor measurements and patient, environmental or acquisition attributes across medical modalities [29].

Together, these studies improve the detection and characterisation of shortcut risk, but identifying a shortcut and changing the representation learned by a classifier are different problems. Dataset auditing can reveal an undesirable association without specifying how pathology-relevant evidence should enter the predictor; post-hoc attribution can reveal where a trained predictor responds without constraining the representation from which that response arose. This leaves an intervention problem: whether pathology-related spatial preference can be introduced during representation learning without requiring tumour-mask supervision, and whether its effect can be distinguished from classification accuracy itself. PFD–GSTE is studied as such an intervention. The present experiment asks whether guidance changes the observed evidence pathway; it does not infer causal removal of shortcut learning from saliency alone.

2.2 CNN, transformer and hybrid architectures

Residual networks improve deep optimisation through identity mappings and remain widely used as medical-image feature extractors [30]. The Visual Geometry Group network (VGG) carries a comparatively large parameter cost, densely connected convolutional networks (DenseNet) promote feature reuse, and EfficientNet couples network depth, width and resolution through compound scaling [31–33]. ResNet50V2 is used here for three implementation properties: pretrained transfer, a 7×7 terminal spatial representation at 224×224 input resolution, and convolutional activations that remain accessible to Grad-CAM++. The implementation instantiates the Big Transfer (BiT-M) `resnetv2_50x1_bitm` checkpoint from `timm`, fixing both the ResNetV2 architecture and pretrained-weight family rather than using ResNet50V2 as an architectural label alone [34, 35].

Transformers provide a complementary representation of spatial context. ViT tokenises an image and models token interactions through self-attention; the data-efficient image Transformer (DeiT) improves data efficiency, Swin Transformer constrains attention through shifted local windows, and the Convolutional Vision Transformer (CvT) introduces convolutional structure into transformer processing [15, 36–38]. For brain MRI, Krishnan et al. introduce the Rotation-Invariant Vision Transformer (RViT), which incorporates $0°$, $90°$, $180°$ and $270°$ views before transformer encoding [16]. The modified ResNet50V2 classifier of Sarada et al. provides the direct CNN-side architectural motivation for this project [39]. Subsequent brain-MRI systems combine ResNet50V2 with Swin Transformer, investigate a distinct Residual Vision Transformer (ResViT) under self-supervised learning, or fuse convolutional and transformer representations under other objectives [40–42].

Recent work has also raised the empirical standard beyond a single high-accuracy internal split. On the same 7,023-slice four-class benchmark family, Zakavi et al. [43] report 99.01% internal test accuracy and a five-fold mean of $99.33 \pm 0.17\%$, followed by 97.55% accuracy on the Brain Tumor Image Segmentation and Classification (BRISC) 2025 dataset without fine-tuning. This directly tests whether high internal classification performance persists on a dataset with different imaging characteristics. The study nevertheless remains slice-level and lacks patient-level stratification, and its authors identify patient-level, multimodal and prospective evaluation as necessary for clinical validation. Houssein et al. [44] approach a different limitation, combining fine-tuned ResNet50 classification with SHapley Additive exPlanations (SHAP) [45] and reporting 99.30% test accuracy on a four-class dataset; their analysis improves access to pixel-level feature contributions but remains based on a comparatively small two-dimensional dataset. Shahzad et al. [46] similarly combine a Swin-based classifier with Gradient-weighted Class Activation Mapping (Grad-CAM) on a multi-source MRI collection, reporting 98% accuracy and a Matthews correlation coefficient (MCC) of 0.9576.

Together, these results show that very high slice-level classification accuracy is no longer sufficient to define the unresolved problem. External testing addresses distributional transfer; SHAP and Grad-CAM address post-hoc interrogation of a trained predictor. Neither property, by itself, requires the representation used for prediction to privilege pathology-related spatial evidence during learning. The question examined here is therefore not whether another CNN–Transformer combination can raise an already near-saturated internal accuracy. It is whether a pathology-guided intervention can alter where evidence enters a hybrid representation, and whether the resulting spatial behaviour agrees with or diverges from classification performance.

Late fusion provides a controlled setting for this question. The CNN and transformer descriptors are projected independently to 256 dimensions and concatenated only at the classification head. Hybrid A leaves the CNN descriptor ungated while applying PFD–GSTE to the CNN-derived transformer-token pathway. Hybrid B applies spatial guidance to both the CNN descriptor and the raw-image patch pathway. The two architectures therefore examine distinct intervention sites rather than treating CNN–Transformer fusion itself as the contribution.

2.3 Pathology guidance, explainability and uncertainty

Prior work reaches pathology-relevant representations through different forms of supervision and architectural constraint. Pathology-disentanglement methods explicitly seek to separate disease-related and non-disease informa-

Table 1: Closest architectural and evaluation reference points. Reported values are contextual rather than directly ranked because dataset composition, split construction and validation protocols differ.

Study	Model and task	Reported result	Relation to the present study
Krishnan et al. [16]	Rotation-invariant ViT; binary glioma/no-tumour	98.6% accuracy	Defines the four-rotation transformer construction adopted here; no ResNet50V2 fusion or four-class guidance experiment
Sarada et al. [39]	Modified ResNet50V2; four classes	96.33% accuracy	Direct ResNet50V2 inspiration; CNN-only classification without transformer-side pathology guidance
Al Bataineh et al. [40]	Swin Transformer + ResNet50V2; four classes	96.8% accuracy	Establishes a ResNet50V2–Transformer brain-MRI hybrid using a different transformer and no PFD–GSTE intervention
Karagoz et al. [41]	ResViT self-supervised hybrid; four classes	98.47% accuracy	Studies a different Residual Vision Transformer and training objective; no coupled spatial-guidance mechanism
Zakavi et al. [43]	Swin-Tiny; internal and external four-class evaluation	99.01% internal; 97.55% external	Strengthens cross-dataset validation on the same benchmark family; does not test training-time pathology-directed evidence
Houssein et al. [44]	Fine-tuned ResNet50 + SHAP; four classes	99.30% test accuracy	Combines high classification performance with post-hoc pixel attribution; guidance is not part of representation learning
Shahzad et al. [46]	Tumor-Swin Transformer + Grad-CAM	98.0% accuracy; MCC 0.9576	Combines transformer classification and local explanation on a multi-source dataset; no guided-versus-unguided representation intervention
Present work	ResNet50V2–RViT with PFD–GSTE A/B and family-specific controls	98.52–99.22% accuracy	Separates classification performance from the spatial behaviour induced by two mask-free pathology-guidance intervention sites

tion, while tumour-region augmentation exposes a classifier to an identified lesion region [47, 48]. The Dynamic Vision Transformer (DynamicViT) and Token Merging instead alter transformer computation through learned token importance or redundancy and are not pathology-supervised mechanisms [49, 50]. PFD–GSTE differs in where spatial guidance is introduced and how it is supervised. PFD predicts a soft spatial preference from the deepest ResNet50V2 representation under image-level supervision; GSTE transfers that preference to the transformer-facing representation. No verified tumour mask supervises the gate. Consequently, the gate is neither a segmentation target nor evidence of formal pathology–background factorisation.

Explanation is likewise distinct from guidance. Grad-CAM++ estimates class-associated spatial evidence from convolutional activations and gradients [51, 52]. Attention Rollout instead propagates self-attention through successive transformer layers [53]. The two methods therefore interrogate different representation pathways. In the present implementation, Grad-CAM++ is conditioned on the target-class logit and applied to the spatial CNN representation, whereas Attention Rollout aggregates transformer information flow by averaging attention over heads, incorporating residual self-connections, composing the resulting matrices across layers and, because the RViT encoder has no class token, averaging over query tokens to obtain spatial token relevance. Attention Rollout is consequently not a class-specific attribution of the final fused prediction. Its spatial resolution also follows the transformer token grid, which is coarser for the 49-token Hybrid A pathway than for the image-patch Hybrid B pathway. Neither method establishes that a highlighted region causally determined the final prediction [54–56]; they are used here as complementary branch-specific diagnostic views rather than interchangeable localisation measures.

Recent medical artificial intelligence (AI) research extends explanation beyond individual attribution maps. Xie et al. [57] introduce class-association manifold learning, extracting class-associated decision structure from trained medical AI models and generating counterfactual examples that expose class-level decision rules. This advances the question from "where did this model respond on one image?" towards "what common structure defines its learned decision rule?" Yet explanation remains analytically distinct from intervention: extracting a decision rule after training does not constrain which evidence entered the representation while the classifier was being learned.

PFD–GSTE investigates the complementary training-time question. Rather than applying an explanatory method only after optimisation, it inserts a learned spatial preference into the representation pathway itself. This makes it possible to ask whether stronger pathology-directed guidance changes observed tumour-centred evidence even when classification accuracy does not improve. The distinction is central to the experimental design: classification, Grad-CAM++, Attention Rollout and failure-case uncertainty are reported separately because agreement between them is an empirical question rather than an architectural assumption.

MC dropout addresses another distinct quantity. Repeated stochastic forward passes provide a practical approximation to predictive variation [58], but confidence and calibration are not equivalent. Distribution-level calibration requires quantities such as negative log-likelihood, Brier score and expected calibration error [59, 60]. The 20-pass experiment reported here is therefore a failure-case probe: it tests whether stochastic variation exposes one shared confidently incorrect subtype prediction. It is not used to rank the four models by calibrated uncertainty.

2.4 MRI-specific foundation representations

MRI representation learning is also moving from task-specific two-dimensional classifiers towards reusable three-dimensional representations. Triad is pretrained on 129,302 three-dimensional MRI volumes from 19,104 patients and evaluates transfer across segmentation, classification and registration [61]. Its results support a broader principle: modality-specific pretraining can provide representations that transfer across MRI tasks rather than being optimised for a single classification dataset. Decipher-MR extends this direction through self-supervised visual learning and radiology-report supervision over more than 200,000 MRI series from more than 22,000 studies, spanning heterogeneous anatomical regions, sequences, pathologies and scanner manufacturers [62]. It evaluates the resulting representation across disease classification, demographic prediction, anatomical localisation and cross-modal retrieval.

These models address weaknesses that the present dataset cannot: volumetric context, acquisition heterogeneity and representation transfer across tasks and clinical domains. Their success, however, does not make representation breadth equivalent to pathology-specific evidence for an individual downstream decision. A representation may transfer well while a downstream classifier still exploits an unintended predictive feature; conversely, suppressing non-lesion information too aggressively may remove context that is genuinely discriminative between tumour subtypes. Scale and evidence specificity are therefore related but different problems.

The present study isolates the latter question in a substantially smaller experimental regime. Hybrid A preserves an ungated CNN summary while imposing PFD–GSTE on the transformer-facing feature representation. Hybrid B extends the same spatial preference to both the CNN descriptor and image-patch pathway. Their family-specific controls test whether changing the intervention site changes classification and observed spatial evidence. This experiment cannot establish that the mechanism will transfer to patient-level, multi-sequence or foundation-model representations; rather, those settings provide the appropriate next test of whether the observed behaviour persists beyond the present slice-level benchmark.

Contemporary medical imaging has therefore advanced along three complementary axes: increasingly strong internal and external classification, increasingly sophisticated methods for auditing or explaining model behaviour, and increasingly transferable MRI representations. A gap remains between these capabilities. Predictive success does not establish pathology-relevant evidence; detecting a shortcut does not remove it; explaining a trained decision does not constrain the representation from which that decision arose; and broader pretraining does not by itself determine which feature a downstream classifier will exploit.

This study addresses a narrower question: can pathology-related spatial preference be inserted directly into a CNN–Transformer representation, without tumour-mask supervision, such that its effect on classification and observed spatial evidence can be examined separately? PFD–GSTE is introduced as a mechanism for this purpose. The guided and unguided family-specific controls do not assume that stronger guidance must improve classification. Instead, they test whether changing where spatial preference enters the representation alters classification, tumour-centred evidence or failure structure. Under this formulation, disagreement between classification and observed spatial evidence is itself informative: it exposes the distinction between predicting the correct class and relying on evidence that is visually associated with the intended pathology.

3 Method

3.1 Dataset curation and split construction

Every file in the Kaggle Training and Testing directories of the Nickparvar brain-tumour MRI dataset [25] is entered into a manifest before modelling. A SHA-1 digest [63] is computed from the file bytes. If the same digest occurs in both source directories, the Testing copy is retained and the Training copy is removed. Byte-level hashing gives an objective and reproducible test for exact copies before any split is created. Retaining the Testing copy preserves the source test directory while preventing the same file from also entering training. The procedure gives

$$7{,}023 \text{ raw files} - 297 \text{ duplicates} = 6{,}726 \text{ retained images}. \tag{1}$$

Images are corrected for Exchangeable Image File Format (EXIF) orientation, tightly cropped with intensity threshold 5 and margin 10, resized to 224×224 with Lanczos resampling, and converted to RGB. Cropping reduces large empty borders while the ten-pixel margin avoids cutting directly against visible anatomy. The 224×224 size satisfies the backbone input and gives a 14×14 grid of non-overlapping 16×16 patches. RGB conversion gives one fixed three-channel input contract for every source image. The cleaned Kaggle Training directory is divided into stratified training and validation sets (80/20, seed 42) using scikit-learn [64]; the cleaned Kaggle Testing directory remains held out so that checkpoint selection never uses test labels.

Dataset snapshot. All dataset counts, SHA-1 hashes and split statistics reported in this study refer to the 7,023-file snapshot used when the experiments were conducted. The externally hosted dataset is versioned independently and may subsequently change; later revisions of the hosting data card are therefore not interchangeable with the experimental snapshot documented here.

3.2 PFD spatial gate

Let B denote batch size, $\mathbb{R}$ the real numbers and $x \in \mathbb{R}^{B\times3\times224\times224}$ an input batch. The backbone mapping $\mathcal{B}_\theta$, with parameters θ, produces the deepest ResNet50V2 feature map $F = \mathcal{B}_\theta(x) \in \mathbb{R}^{B\times C\times7\times7}$, where $C = 2048$ is the channel count. Let $*_{1\times1}$ denote a learned 1×1 convolution, σ the logistic sigmoid and $\odot$ element-wise multiplication.

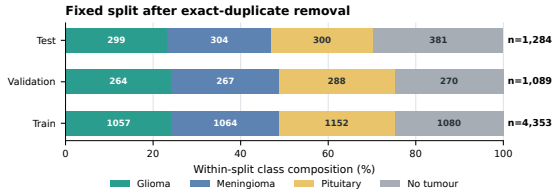


Figure 2: Fixed split composition after exact-duplicate removal. Counts inside each segment give the number of images in each class.

With kernel $W_m \in \mathbb{R}^{1\times C\times 1\times 1}$ and scalar bias b_m, the PFD gate M and gated feature map F^g are

$$M = \sigma(W_m *_{1\times 1} F + b_m), \qquad F^g = F \odot M. \tag{2}$$

$M \in [0,1]^{B\times 1\times 7\times 7}$ is broadcast across the C channels. No segmentation mask supervises M; the gate is learned jointly through the four-class objective.

The term *Pathology-Focused Disentanglement* (PFD) is retained as the experiment-specific name of this module. In the present implementation, disentanglement is operational rather than a formal latent-variable factorisation: PFD learns a one-channel multiplicative spatial preference map that differentially weights feature locations under image-level supervision. It does not impose statistical independence, mutual-information minimisation, adversarial separation or an explicit pathology–background factorisation. Instead, PFD provides the learned spatial preference from which the downstream GSTE operation constructs representation-level guidance.

3.3 Two guidance strengths and two ablations

The experiment tests spatial guidance at two intervention sites. In Hybrid A, PFD–GSTE changes only the transformer-facing feature tokens; the CNN descriptor remains ungated. In Hybrid B, the gate changes both the CNN descriptor and the raw-image patch tokens, and GSTE-B can reduce the token grid when the gate is concentrated. These variants instantiate two intervention patterns: feature-token guidance in Hybrid A and joint image-token and CNN guidance in Hybrid B.

Each guided model has a family-specific control. Ablation A rotates and encodes the ungated 7×7 CNN feature map without PFD gating or GSTE-A token reweighting. Ablation B retains the ungated CNN descriptor and the complete 14×14 image-token grid, removing PFD-B, GSTE-B token weighting and the associated adaptive grid reduction.

These controls are mechanism-level rather than component-wise ablations. GSTE is defined in the present architecture as consuming the spatial preference produced by PFD; the principal experimental unit is therefore the implemented PFD–GSTE mechanism within each family. The contrasts Hybrid A versus Ablation A and Hybrid B versus Ablation B estimate the effect of introducing or removing that complete coupled mechanism while preserving the corresponding family architecture. They do not identify the separate causal contribution of PFD, GSTE or, in Hybrid B, adaptive grid reduction. A factorial study containing PFD-only, token-guidance-only, complete PFD–GSTE and unguided controls, together with a grid-reduction-only Hybrid B control, is reserved for subsequent evaluation.

Cross-family comparisons are treated separately because Hybrid A and Hybrid B additionally differ in token source, CNN-descriptor intervention, internal attention projection width and data-loader behaviour; they are not interpreted as a single-factor ablation of guidance strength.

3.4 Hybrid A: feature-token guidance

Let global average pooling (GAP) and the rectified linear unit (ReLU) be denoted by GAP and ReLU, respectively. With learned projection parameters W_c^A and b_c^A, Hybrid A forms the ungated CNN descriptor z_{cnn}^A as

$$z_{\mathrm{cnn}}^A = \mathrm{ReLU}\left(W_c^A\,\mathrm{GAP}(F) + b_c^A\right). \tag{3}$$

For a rotation index $r \in \{0,1,2,3\}$, let $\mathcal{R}_r$ rotate its argument by $90r$ degrees and let vec flatten a spatial grid into a token sequence. The token index is i, and the numerical stabiliser is $\varepsilon = 10^{-6}$. Learned parameters W_e^A and b_e^A project the rotated feature map to 142 channels; $P_{7\times 7} \in \mathbb{R}^{49\times 142}$ is the learned positional embedding and $\rho_r \in \mathbb{R}^{142}$ the learned embedding for rotation r. The projected tokens E_r^A, mean-normalised token weights a_r^A and guided sequence T_r^A are

$$E_r^A = \mathrm{vec}\left(W_e^A *_{1\times 1} \mathcal{R}_r(F^g) + b_e^A\right) \in \mathbb{R}^{B\times 49\times 142}, \tag{4}$$

$$a_r^A = \frac{\mathrm{vec}(\mathcal{R}_r(M))}{\mathrm{mean}_i\,\mathrm{vec}(\mathcal{R}_r(M))_i + \varepsilon}, \tag{5}$$

$$T_r^A = E_r^A \odot a_r^A + P_{7\times 7} + \rho_r. \tag{6}$$

Coupled two-stage guidance Hybrid A applies the learned PFD spatial preference at two successive representations of the transformer-facing pathway. First, PFD-A forms the gated CNN representation $F^g = F \odot M$ before feature projection. Second, after the rotated gated feature map is projected into the 49-token sequence E_r^A, GSTE-A derives the mean-normalised weights a_r^A from the same spatial gate and applies $E_r^A \odot a_r^A$. The two multiplications therefore act at different representation levels: PFD-A modifies the convolutional feature field, whereas GSTE-A transfers its relative spatial preference to the projected token field consumed by RViT. This serial construction is intentional in the implemented method rather than a second independent pathology estimator. The late-fusion CNN descriptor remains computed from ungated F and consequently bypasses both guidance operations.

Mean normalisation keeps the average token scale close to one, so the gate changes relative spatial emphasis rather than uniformly shrinking the sequence. GSTE-A changes token magnitude but keeps the 7×7 grid. The rotation-averaged sequence $\overline{T}^A$ supplied to the encoder is

$$\overline{T}^A = \frac{1}{4}\sum_{r=0}^{3} T_r^A. \tag{7}$$

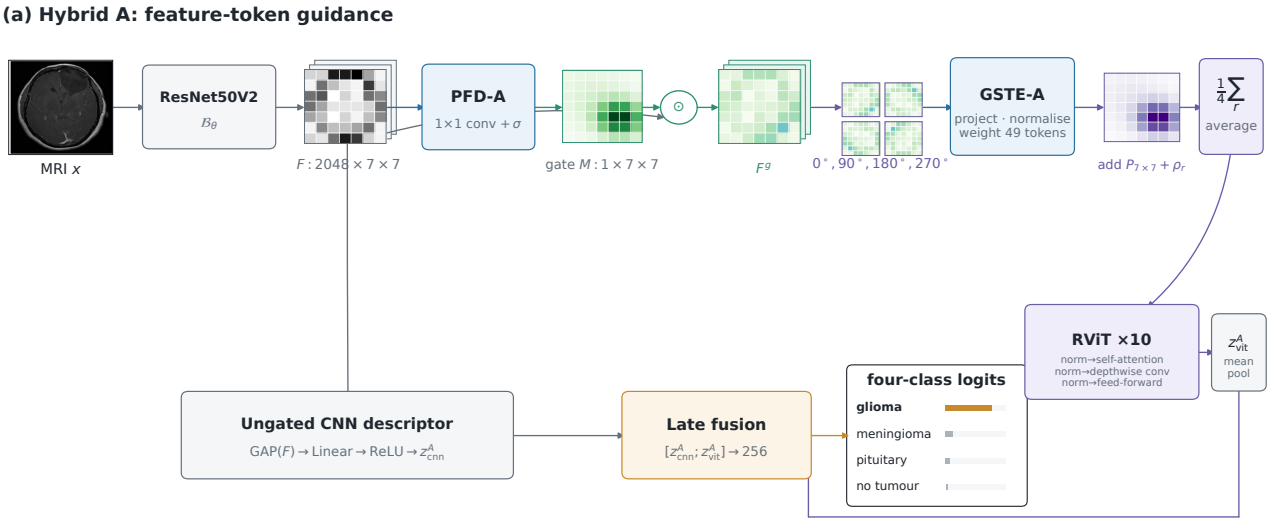


Figure 3: **Hybrid A: feature-token guidance.** The input slice x is encoded by ResNet50V2 as $F \in \mathbb{R}^{B\times2048\times7\times7}$. PFD-A predicts M and forms F^g. Four aligned rotations of F^g and M are projected to 49×142 feature-token sequences. GSTE-A mean-normalises the aligned gate, weights the 49 tokens, adds the positional embedding $P_{7\times7}$ and rotation embedding ρ_r, and averages the four views before ten RViT blocks. The lower branch pools the ungated F; therefore, the CNN descriptor bypasses PFD-A. Late fusion combines the 256-dimensional CNN and transformer descriptors to produce the four-class logits. Thus, the transformer-facing branch receives coupled feature-domain and token-domain guidance from the same learned spatial preference, whereas the late-fusion CNN descriptor remains ungated.

3.5 Hybrid B: patch-token guidance and bounded grid reduction

Let batch normalisation (BN) [65] be denoted by BN, and let Dropout denote standard dropout [66]. With learned projection parameters W_c^B and b_c^B, Hybrid B forms the gated CNN descriptor z_{cnn}^B as

$$z_{\text{cnn}}^B = \text{ReLU}\big[\text{BN}\big(W_c^B \,\text{Dropout}(\text{GAP}(F^g)) + b_c^B\big)\big]. \tag{8}$$

The image is divided into non-overlapping 16×16 patches. For each rotation, a strided convolution forms $14 \times 14 = 196$ tokens $E_r^B \in \mathbb{R}^{B\times196\times142}$. Let up denote bilinear upsampling to image resolution, and let adaptive average pooling (AAP) to an $h \times w$ grid be denoted by $\text{AAP}_{h\times w}$; indices i and j address grid locations. The mean-normalised patch gate $A_r \in \mathbb{R}^{B\times1\times14\times14}$ is

$$A_r = \frac{\text{AAP}_{14\times14}(\mathcal{R}_r(\text{up}(M)))}{\text{mean}_{i,j}\,\text{AAP}_{14\times14}(\mathcal{R}_r(\text{up}(M)))_{ij} + \varepsilon}. \tag{9}$$

GSTE-B selects one batch-consistent grid side $s \in [7, 14]$. The concentration statistic q is the batch mean of the spatial sample standard deviation of the unrotated gate A_0. A nearly uniform gate provides no basis for reducing the grid, whereas a concentrated gate permits greater reduction. The lower and upper concentration thresholds are $q_f = 0.05$ and $q_c = 0.30$, and the maximum fractional side reduction is $\gamma = 0.50$. The normalised concentration u and selected side s are defined below, where $\text{clip}(v, l, h)$ truncates v to $[l, h]$ and round returns the nearest integer:

$$u = \text{clip}\left(\frac{q - q_f}{q_c - q_f}, 0, 1\right), \tag{10}$$

$$s = \begin{cases} 14, & q \le q_f, \\ \text{clip}(\text{round}(14(1-\gamma u)), 7, 14), & q > q_f. \end{cases} \tag{11}$$

The side is bounded below by seven to prevent the patch pathway from collapsing to a very coarse representation. One side length is used for the whole batch, preserving a common tensor shape during training. Reshaping E_r^B to the patch grid $X_r \in \mathbb{R}^{B\times142\times14\times14}$, GSTE-B computes the guided grid $\widehat{X}_r \in \mathbb{R}^{B\times142\times s\times s}$ as

$$\widehat{X}_r = \begin{cases} X_r \odot A_r, & s = 14, \\ \dfrac{\text{AAP}_{s\times s}(X_r \odot A_r)}{\text{AAP}_{s\times s}(A_r) + \varepsilon}, & s < 14. \end{cases} \tag{12}$$

The denominator makes the reduced cells weighted averages rather than unnormalised sums. Let $\widetilde{P}_s \in \mathbb{R}^{s^2\times142}$ denote the learned 14×14 positional embedding interpolated to side s. Flattening $\widehat{X}_r$ and adding the positional and rotation embeddings gives $T_r^B = \text{vec}(\widehat{X}_r) + \widetilde{P}_s + \rho_r \in \mathbb{R}^{B\times s^2\times142}$. The encoder receives the rotation average $\overline{T}^B = \frac{1}{4}\sum_{r=0}^{3} T_r^B$. Averaging occurs before the transformer so that one orientation-balanced sequence, rather than four independent predictions, enters the common encoder.

Hybrid B is intentionally a broader guidance intervention than Hybrid A. Whereas Hybrid A preserves an ungated

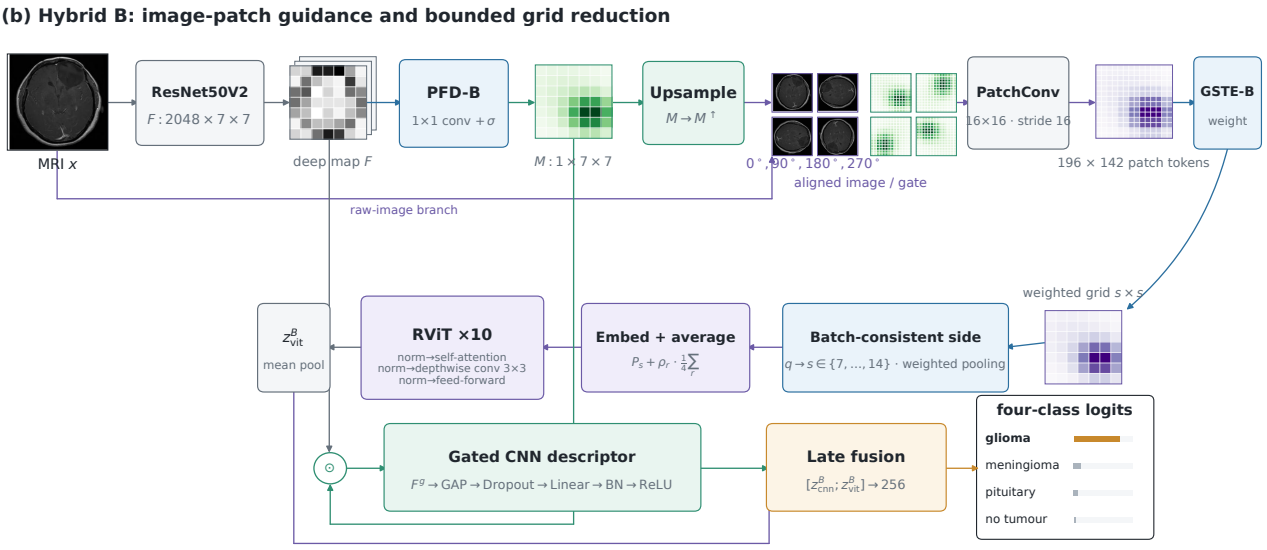


Figure 4: **Hybrid B: image-patch guidance and bounded grid reduction.** ResNet50V2 produces F, and PFD-B predicts the 7×7 gate M. The input slice and upsampled gate $M^{\uparrow}$ are rotated together. A 16×16 convolution with stride 16 produces 196×142 patch-token sequences. GSTE-B applies the aligned, mean-normalised gate and selects one batch-consistent side $s\in\{7,\dots,14\}$; when $s<14$, weighted adaptive pooling reduces the token grid while preserving one sequence length across the batch. Position and rotation embeddings are added before the four views are averaged and passed through ten RViT blocks. The lower branch pools F^g; therefore, PFD-B also controls the CNN descriptor before late fusion.

CNN descriptor and confines the coupled PFD–GSTE mechanism to the transformer-facing feature pathway, Hybrid B allows the learned PFD preference to affect both branches of the fused classifier: F^g supplies the pooled CNN descriptor and the same spatial preference guides the image-patch sequence through GSTE-B. The concentration-dependent grid operation additionally tests whether a spatially informative gate can concentrate the token representation without introducing a separate lesion detector. Hybrid B is therefore not interpreted as a single-factor modification of Hybrid A; it is the stronger of two deliberately different intervention sites whose controlled comparisons are made against their own family-specific ablations.

3.6 RViT encoder and fusion

Both variants share the same external RViT depth, embedding width, head count and feed-forward width, while their internal attention projection widths differ as specified below. Let H_0 be the rotation-averaged input sequence $\overline{T}^A$ or $\overline{T}^B$, and let $\ell\in\{1,\dots,10\}$ index encoder blocks. Layer normalisation (LN), multi-head self-attention (MHSA), depthwise convolution (DWConv) and a multilayer perceptron (MLP) are the four named block operators inherited from the adopted transformer/RViT design [14, 16]. The operators grid and vec reshape a token sequence to its square spatial grid and back, and $\mathcal{D}_{0.1}$ denotes dropout with probability 0.1. Intermediate residual states are denoted by U_ℓ and V_ℓ, and the output of block ℓ by H_ℓ. The code implements

$$U_\ell = H_{\ell-1} + \mathrm{MHSA}(\mathrm{LN}(H_{\ell-1})), \tag{13}$$

$$V_\ell = U_\ell + \mathcal{D}_{0.1}(\mathrm{vec}[\mathrm{DWConv}_{3\times3}(\mathrm{grid}(\mathrm{LN}(U_\ell)))]), \tag{14}$$

$$H_\ell = V_\ell + \mathrm{MLP}(\mathrm{LN}(V_\ell)). \tag{15}$$

The multi-head attention and multilayer-perceptron operators include their code-specified 0.1 dropout, and the depthwise-convolution update is followed by the same dropout rate. The external embedding dimension is 142 with ten heads, following the adopted RViT configuration [16]. Because 142 is not divisible by ten, Hybrid A uses head width $\lfloor 142/10\rfloor = 14$ and an internal attention width of 140; Hybrid B uses $\lceil 142/10\rceil = 15$ and an internal width of 150. Each attention output is projected back to 142 dimensions. After the tenth block, the encoder applies a final layer normalisation, $\widetilde{H} = \mathrm{LN}(H_{10})$.

With learned projection parameters W_v and b_v, the final transformer descriptor $z_{\mathrm{vit}}\in\mathbb{R}^{B\times256}$ is obtained by mean-pooling the normalised sequence:

$$z_{\mathrm{vit}} = \mathrm{ReLU}\left(W_v\,\mathrm{mean}_i(\widetilde{H}_i) + b_v\right). \tag{16}$$

Late fusion keeps the local CNN summary z_{cnn} and contextual transformer summary z_{vit} separate until both have the same 256-dimensional width. Let W_f, b_f parameterise the fusion projection, let W_o, b_o parameterise the output projection, and let $[a;b]$ denote vector concatenation. The fused hidden vector h and four-class logit vector $\widehat{y}\in\mathbb{R}^{B\times4}$

are

$$h = \mathrm{ReLU}(W_f[z_{\mathrm{cnn}}; z_{\mathrm{vit}}] + b_f)\,, \tag{17}$$

$$\widehat{y} = W_o\,\mathrm{Dropout}(h) + b_o. \tag{18}$$

The complete code-derived forward procedures are given in Algorithms 1 and 2.

4 Training and Evaluation

The models are implemented in PyTorch [67] and trained for at most 100 epochs with the Adam optimiser [68] using decoupled weight decay (AdamW) [69]. The pretrained CNN uses learning rate 10^{-4}; the newly initialised transformer and fusion parameters use 5×10^{-4}. The lower CNN rate limits early changes to transferred features. The CNN is frozen for epochs 1–5, during which only the new guidance, transformer and fusion layers are updated; joint fine-tuning begins at epoch 6. Weight decay is 0.01, excluding biases, normalisation parameters and one-dimensional parameters. Cross-entropy uses label smoothing 0.05 [70]. Cosine annealing reduces the learning rate to 10^{-6} [71], and gradient clipping bounds the norm at 1.0 [72].

The batch size is 32, early-stopping patience is ten, and the checkpoint with the highest validation macro-F1 is retained. Macro-F1 averages the four class-wise F1 scores equally, whereas accuracy is sample-weighted. Training augmentation comprises rotation in $[-15°, 15°]$, horizontal flip with probability 0.5, translation up to 5%, and Gaussian noise with standard deviation 0.02 and probability 0.5. These stochastic transforms regularise orientation, position and intensity at the input. RViT performs a separate operation inside each forward pass by averaging tokens from four exact rotations. Validation and test transforms are deterministic.

Each configuration has one training run with seed 42. Hybrid B and Ablation B use `drop_last=True`; the A family retains the final incomplete batch. This difference prevents a controlled numerical comparison between A and B, but does not affect the within-family comparisons. Automatic mixed precision and explicit CNN BN freezing are not used.

Classification is measured on all 1,284 test images using parameter count (Params), accuracy (Acc.), cross-entropy loss, macro-F1, weighted F1 score (W-F1), macro-averaged specificity (Spec.), Cohen's kappa (κ) [73], Matthews correlation coefficient (MCC) [74, 75] and exact error counts. The F1 score is the harmonic mean of precision and recall; macro-F1 weights the four classes equally, whereas W-F1 weights them by sample count. Confusion matrices identify the remaining class boundaries. Cohen's κ provides chance-corrected agreement, whereas the multiclass MCC summarises correlation between predicted and true class assignments.

Grad-CAM++ [52] and Attention Rollout [53] are used as complementary branch-specific diagnostics. Grad-CAM++ is computed from the spatial CNN representation and is conditioned on the target-class logit. Attention Rollout uses the attention matrices returned by the transformer encoder: attention is averaged across heads, augmented with the identity, row-normalised and composed across layers; because the encoder contains no class token, the resulting rollout is averaged over query tokens to obtain one relevance value per spatial token. The rollout therefore visualises transformer information flow rather than class-specific attribution of the complete late-fused prediction. Its spatial granularity follows the active token grid, giving a fixed 7×7 map in the A family and an $s \times s$, $7 \leq s \leq 14$, map in the B family before interpolation to image resolution.

The three held-out cases recorded are retained as a fixed qualitative audit: a correctly classified meningioma, a challenging glioma and a pituitary image predicted as meningioma. The same cases are examined for all four models, preventing model-specific case selection. Their qualitative labels were assigned manually by comparing the dominant highlighted region with the visible lesion under a common judgement rule. No tumour masks or blinded clinical annotations were available, and three visual examples cannot establish localisation performance. MC dropout [58] uses 20 forward passes with BN in evaluation mode; dropout changes the probability samples while the learned normalisation statistics remain fixed. Their mean and variance are treated as uncertainty signals, not calibration estimates.

5 Results

5.1 Held-out classification

All four variants exceed 98.5% test accuracy. Ablation B is the strongest classifier, with 99.22% accuracy, 0.9920 macro-F1 and ten errors. Hybrid B records 98.52%, 0.9849 and 19 errors. Within the A family, guidance leaves accuracy unchanged and changes macro-F1 by $+0.0002$. Within the B family, guidance changes accuracy by -0.0070 and macro-F1 by -0.0071.

5.2 Error structure

The remaining errors are concentrated mainly at tumour subtype boundaries. Hybrid A and Ablation A each make 16 errors, but with different directions. Hybrid B makes 19 errors, including five pituitary-to-meningioma errors and seven no-tumour false positives: six glioma and one meningioma prediction. Ablation B makes ten errors and has the most balanced matrix. One pituitary-to-meningioma error is shared by every model. Sellar and suprasellar meningiomas can resemble pituitary adenomas on MRI, which makes this particular subtype confusion clinically plausible [76, 77].

5.3 Explainability and uncertainty

The fixed qualitative audit separates CNN-side class attribution from transformer-side information flow. Under the manual Grad-CAM++ judgements recorded for these three cases, Hybrid B is tumour-centred in all three, Hybrid A in none, and each ablation in one. The corresponding Attention

Table 2: Held-out results on 1,284 images. Bold indicates the best observed value.

Model	Params (10^6)	Acc.	Macro-F1	W-F1	Loss	Spec.	κ	MCC	Errors
Hybrid A	26.68	0.9875	0.9875	0.9875	0.0803	0.9959	0.9833	0.9833	16
Hybrid B	26.58	0.9852	0.9849	0.9852	0.0753	0.9952	0.9802	0.9803	19
Ablation A	26.68	0.9875	0.9873	0.9875	0.0847	0.9959	0.9833	0.9834	16
Ablation B	26.58	**0.9922**	**0.9920**	**0.9922**	**0.0668**	**0.9975**	**0.9896**	**0.9896**	**10**

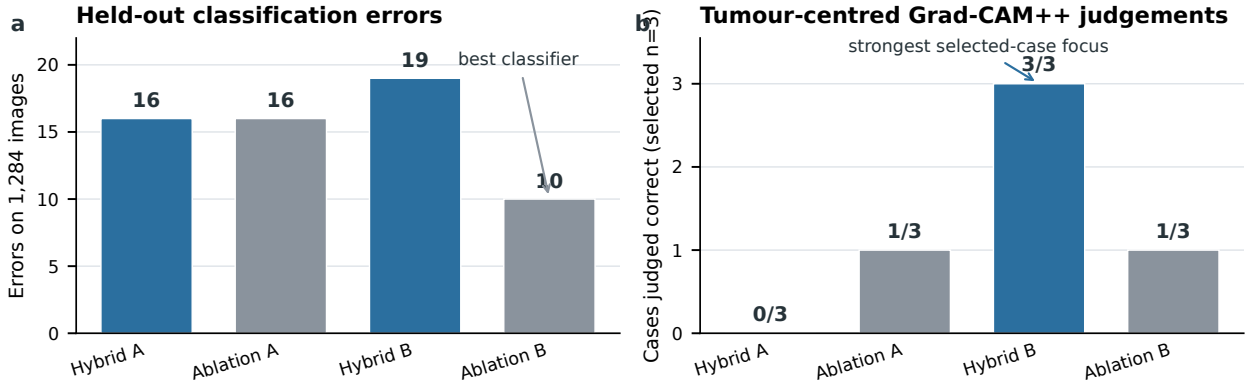


Figure 5: **Classification and qualitative spatial evidence.** (a) Exact classification errors on the 1,284-image held-out test set; Ablation B makes the fewest errors. (b) Number of the three fixed explanation cases judged tumour-centred by Grad-CAM++; Hybrid B is tumour-centred in all three cases. The two panels measure different properties: classification correctness and qualitative spatial evidence.

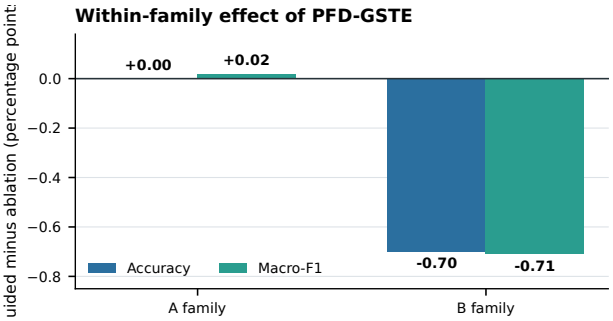


Figure 6: Effect of PFD–GSTE within each family, expressed as guided minus ablation in percentage points.

Rollout maps contain tumour-associated or partially tumour-associated relevance in several cases, but their spatial correspondence with the visible lesion is less precise in this fixed sample. This difference should not be interpreted as failure of the transformer representation. Grad-CAM++ is conditioned on the target-class logit, whereas the implemented Attention Rollout is not class-specific: attention is averaged across heads, residual self-connections are incorporated, the layer-wise attention matrices are composed, and—because the RViT encoder contains no class token—the resulting relevance is averaged over query tokens. Rollout therefore characterises transformer-side information flow rather than attribution of the complete late-fused prediction.

Spatial resolution further depends on the active transformer token grid. Hybrid A represents the transformer-facing feature field with only $7 \times 7 = 49$ spatial tokens, which limits the precision with which its rollout can be projected back to the image. Hybrid B begins from a 14×14 image-patch grid and retains an $s \times s$ representation, with $7 \leq s \leq 14$, after any GSTE-B grid reduction. Its rollout can therefore preserve finer spatial structure than the Hybrid A pathway. These architectural differences motivate interpreting Grad-CAM++ and Attention Rollout as complementary branch-specific diagnostics rather than requiring identical spatial agreement from the two methods.

Beyond these three fixed cases, a broader manual inspection of the Grad-CAM++ and Attention Rollout outputs was conducted during the project. Across the reviewed explanation outputs, approximately 90% of the Hybrid B cases and 70% of the Hybrid A cases were recorded as tumour-focused overall; approximately 30% of the reviewed Hybrid A cases were instead judged to emphasise background or acquisition-related structure despite a correct class prediction. These figures summarise the project-level visual review across the explanation outputs and are not method-specific localisation rates. The review denominator and per-case scoring sheet were not retained, so the proportions are reported only as descriptive historical observations and are excluded from quantitative model comparison.

The fixed-case judgements are likewise qualitative. They were assigned manually by comparing the visible lesion with the explanation overlays, without tumour masks, blinded readers or quantitative overlap measurements. They therefore characterise the observed explanation behaviour of the released models rather than establish localisation accuracy or causal dependence on the highlighted regions.

All four models remain highly confident on the shared error: 0.9692 for Hybrid A, 0.9869 for Hybrid B, 0.9592 for Abla-

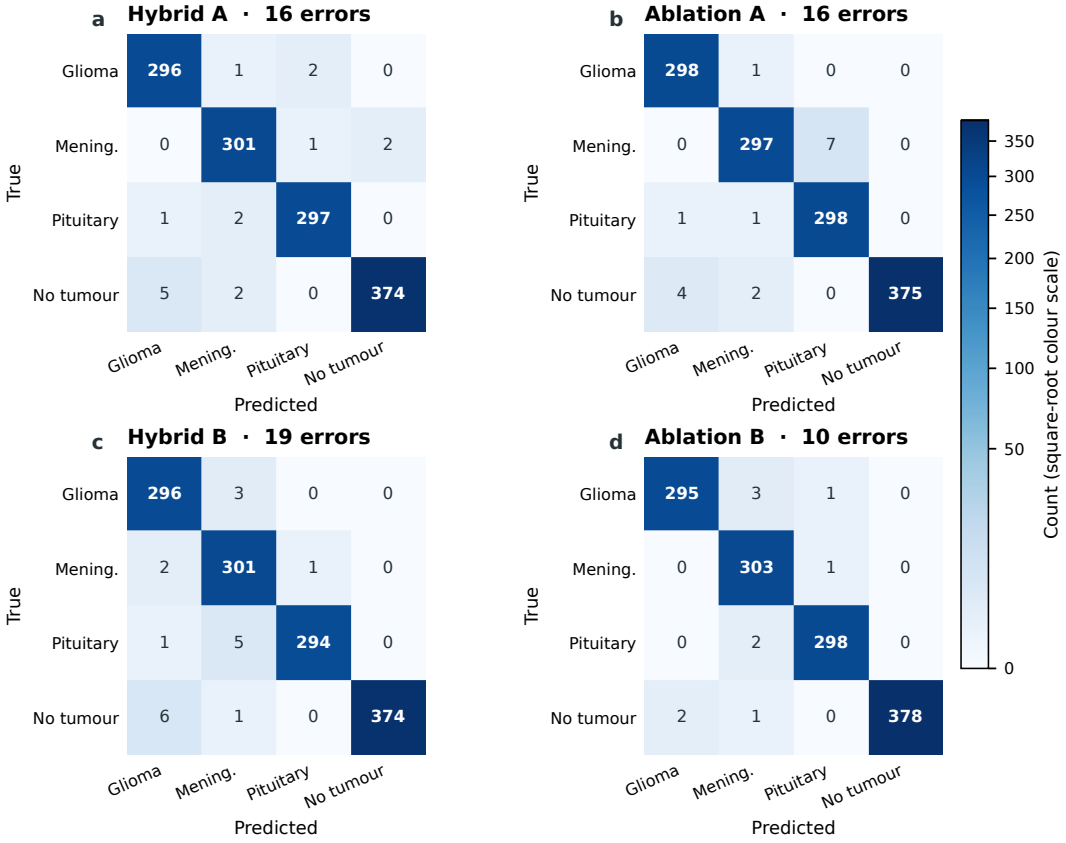


Figure 7: Exact held-out confusion matrices. Rows are true classes and columns are predicted classes.

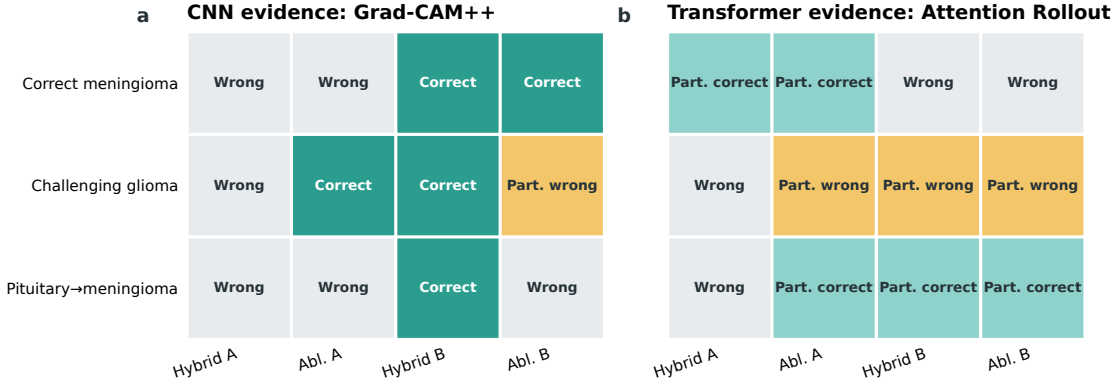


Figure 8: **Branch-specific evidence on three fixed test cases.** (a) Grad-CAM++ examines class-discriminative evidence in the CNN path. (b) Attention Rollout traces information flow through the transformer path. Each cell records the manual visual judgement for the same correctly classified meningioma, challenging glioma and shared pituitary-to-meningioma error. The judgements are qualitative and should not be interpreted as tumour-segmentation scores.

tion A and 0.9632 for Ablation B. MC-dropout confidence does not flag this failure. Hybrid B also produces tumour-centred CNN evidence while assigning the wrong subtype; lesion-centred evidence and correct subtype identification are therefore distinct properties.

6 Discussion and Limitations

Classification and qualitative evidence do not rank the models identically. Ablation B gives the highest macro-F1 (0.9920), whereas Hybrid B shows the strongest tumour-centred Grad-CAM++ behaviour in the fixed manual audit and the highest tumour-focused proportion in the broader descriptive review. Grad-CAM++ and Attention Rollout

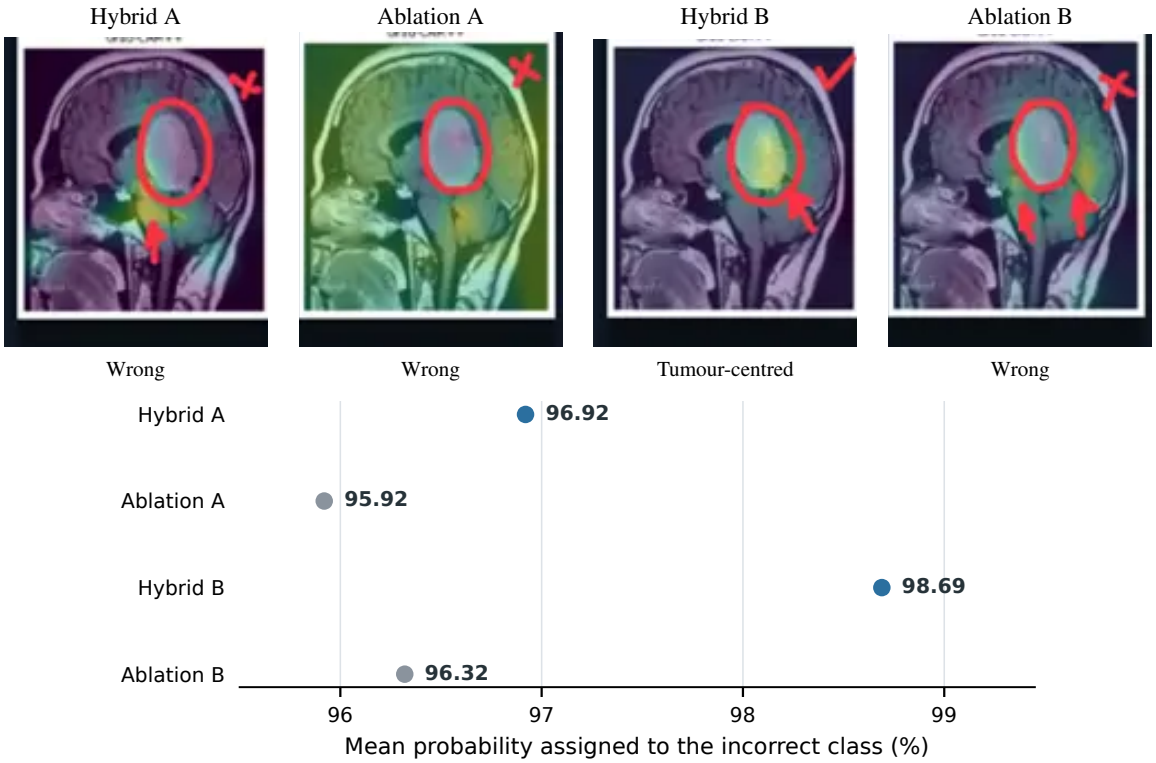


Figure 9: **Confident subtype error.** All four models predict meningioma for a pituitary case. Top: Grad-CAM++ evidence; the red circle marks the visible lesion, which Hybrid B highlights despite assigning the wrong subtype. Bottom: mean probability assigned to the incorrect class over 20 MC-dropout passes.

should not be collapsed into a single localisation measure. Grad-CAM++ is class-conditioned and interrogates the CNN evidence pathway, whereas the implemented Attention Rollout traces information flow within the transformer pathway independently of the target-class logit and does not account for the CNN descriptor or the late-fusion classifier. Its apparent spatial precision additionally depends on token resolution. The more diffuse or partial rollout patterns observed in the fixed cases therefore do not contradict the stronger Hybrid B Grad-CAM++ result; the two methods expose different properties of the hybrid representation.

The broader manual review recorded approximately 90% tumour-focused explanation behaviour for Hybrid B and 70% for Hybrid A. These proportions remain descriptive because the review denominator and per-case scoring sheet were not retained. Neither this broader review nor the three fixed cases uses tumour masks, blinded clinical readers or causal interventions, and neither therefore establishes quantitative localisation performance or complete removal of shortcut learning.

The family-specific results follow the two intervention sites. Hybrid A leaves the CNN descriptor ungated, allowing late fusion to use a local summary that bypasses PFD. Its guided and ablated forms consequently have almost identical classification scores and limited Grad-CAM++ differences. Hybrid B gates the CNN descriptor and the patch-token pathway; its larger Grad-CAM++ change is consistent with this direct intervention. Its lower accuracy admits at least two explanations: the gate may remove context needed for subtype separation, or it may emphasise lesion features that are insufficient for distinguishing tumour subtypes. The present experiment cannot separate these mechanisms.

All four variants achieve macro-F1 of at least 0.9849 on the fixed test split, including both ablations. The ResNet50V2–RViT construction therefore yields high internal classification performance on this split without PFD–GSTE. The guidance modules address a different quantity: where spatial emphasis enters the two branches. The shared pituitary-to-meningioma error further separates these quantities, because every model assigns high confidence to the wrong subtype while Hybrid B retains tumour-centred CNN evidence.

The evidence is bounded by the dataset and protocol. The images are aggregated two-dimensional slices without patient identifiers, sequence labels or an external hospital cohort. SHA-1 removes exact files but not near-duplicates or repeated patients. Each model has one training run, leaving optimisation variance unknown. The three explanation cases have no masks or blinded readers, and MC dropout is not followed by formal calibration. The results are therefore limited to classification and qualitative evidence on this fixed split; they do not establish patient-independent generalisation, calibrated uncertainty or complete removal of shortcut learning.

7 Conclusion

PFD and GSTE introduce spatial guidance into a ResNet50V2–RViT classifier without tumour masks. GSTE-A transfers the PFD gate to 49 CNN feature tokens, whereas GSTE-B transfers it to up to 196 raw-image patch tokens and the CNN descriptor. The four variants are evaluated on one fixed, duplicate-audited split using classification metrics, exact errors, branch-specific explanations and MC dropout.

Every model exceeds the classification target. Ablation B gives the highest macro-F1, while Hybrid B produces the most tumour-centred Grad-CAM++ outputs in the three reported cases. On this benchmark, stronger guidance changes qualitative spatial evidence more than classification performance. Shortcut mitigation cannot therefore be inferred from accuracy alone; the evidence path, error type and confidence behaviour must be examined separately.

8 Future Work

The future study should convert the present mechanism-level comparison into a factorial evaluation of the guidance mechanism. Each family should include an unguided control, PFD-only guidance, token guidance without PFD-derived spatial preference, and the complete PFD–GSTE mechanism; Hybrid B additionally requires a grid-reduction-only control. Data-loader semantics should be identical across configurations, guidance strength should be swept explicitly, and each configuration should be repeated across at least five independent seeds. Mean performance, dispersion and confidence intervals should replace interpretation of single-run sub-percentage differences.

Shortcut mitigation should then be tested directly on patient-independent external cohorts spanning hospitals, scanner manufacturers, acquisition protocols and MRI sequences. Perceptual or representation-level near-duplicate detection should complement the present exact SHA-1 audit. Where tumour masks are available, explanation quality should be quantified using overlap- and pointing-based measures and complemented by causal insertion, deletion, region-occlusion and model-randomisation tests [55]. Uncertainty should be evaluated across complete held-out and external distributions using negative log-likelihood, Brier score, expected calibration error and reliability diagrams [59, 60], together with selective-risk analysis and error- or out-of-distribution-detection metrics. Multi-slice, multi-sequence and three-dimensional models, including contemporary MRI foundation representations, provide the subsequent test of whether PFD–GSTE retains value beyond the present two-dimensional benchmark.

Acknowledgements

I thank Dr Kheng Lee Koay for supervision, patience and clear feedback throughout this research. I also thank the University of Hertfordshire teaching and support teams, my family, and the maintainers of the public dataset and open-source software used in this work.

References

[1] David N. Louis, Arie Perry, Pieter Wesseling, et al. The 2021 who classification of tumors of the central nervous system: A summary. *Neuro-Oncology*, 23(8):1231–1251, 2021. doi:10.1093/neuonc/noab106.

[2] Michael Weller, Martin van den Bent, Matthias Preusser, Emilie Le Rhun, Joerg C. Tonn, Giuseppe Minniti, Martin Bendszus, Carmen Balana, Olivier Chinot, Linda Dirven, et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nature Reviews Clinical Oncology*, 18(3):170–186, 2021. doi:10.1038/s41571-020-00447-z.

[3] Ken K. Y. Ho, Ursula B. Kaiser, Philippe Chanson, et al. Pituitary adenoma or neuroendocrine tumour: The need for an integrated prognostic classification. *Nature Reviews Endocrinology*, 19(11):671–678, 2023. doi:10.1038/s41574-023-00883-8.

[4] Javier E. Villanueva-Meyer, Marc C. Mabray, and Soonmee Cha. Current clinical brain tumor imaging. *Neurosurgery*, 81(3):397–415, 2017. doi:10.1093/neuros/nyx103.

[5] Ali Gholipour, Nasser Kehtarnavaz, Richard Briggs, Michael Devous, and Kadharbatcha Gopinath. Brain functional localization: A survey of image registration techniques. *IEEE Transactions on Medical Imaging*, 26(4):427–451, 2007. doi:10.1109/TMI.2007.892508.

[6] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi:10.1038/s42256-020-00257-z.

[7] Michael Roberts, Derek Driggs, Matthew Thorpe, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3:199–217, 2021. doi:10.1038/s42256-021-00307-0.

[8] Min Lin, Ning Weng, Kasper Mikolaj, Zaid Bashir, Mikkel B. S. Svendsen, Martin G. Tolsgaard, Anders N. Christensen, and Aasa Feragen. Shortcut learning in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 623–633, 2024. doi:10.1007/978-3-031-72111-3_59.

[9] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future.

npj Digital Medicine, 5:48, 2022. doi:10.1038/s41746-022-00592-y.

[10] Pouria Rouzrokh, Bardia Khosravi, Shabnam Faghani, Marzieh Moassefi, Delaram Vafaei Sadr, Mehdi Golbabaee, Yashaswini Sreekumar, Alexandros Zagouras, Garry Choy, and Bradley J. Erickson. Mitigating bias in radiology machine learning: 1. data handling. *Radiology: Artificial Intelligence*, 4(5):e210290, 2022. doi:10.1148/ryai.210290.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi:10.1038/nature14539.

[12] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. doi:10.1109/TMI.2016.2528162.

[13] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. doi:10.1109/TMI.2016.2535302.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. doi:10.48550/arXiv.1706.03762.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[16] P. T. Krishnan, P. Krishnadoss, M. Khandelwal, D. Gupta, A. Nihaal, and T. S. Kumar. Enhancing brain tumor detection in MRI with a rotation invariant vision transformer. *Frontiers in Neuroinformatics*, 18:1414925, 2024. doi:10.3389/fninf.2024.1414925.

[17] Syed Ali Yazdan, Rashid Ahmad, Naeem Iqbal, Atif Rizwan, Anam Nawaz Khan, and Do-Hyeun Kim. An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD. *Tomography*, 8(4):1905–1927, 2022. doi:10.3390/tomography8040161.

[18] Ilaria Maggio, Enrico Franceschi, Vincenzo Di Nunno, and Alba A. Brandes. Meningioma: Not always a benign tumor. a review of advances in the treatment of meningiomas. *CNS Oncology*, 10(2):CNS72, 2021. doi:10.2217/cns-2021-0003.

[19] Michael Amoo, John Henry, Michael Farrell, and Mohsen Javadpour. Meningioma in the elderly. *Neuro-Oncology Advances*, 5(Supplement 1):i13–i25, 2023. doi:10.1093/noajnl/vdac107.

[20] Ian F. Pollack, Sameer Agnihotri, and Alberto Broniscer. Childhood brain tumors: Current management, biological insights, and future directions. *Journal of Neurosurgery: Pediatrics*, 23(3):261–273, 2019. doi:10.3171/2018.10.PEDS18377.

[21] Gabriela S. P. Giraldo, Liron Singer, Tina Cao, et al. Differential diagnosis of tumor-like brain lesions. *Neurology: Clinical Practice*, 13(5):e200182, 2023. doi:10.1212/CPJ.0000000000200182.

[22] Fabrice Bonneville, H. Rolf J"ager, and James G. Smirniotopoulos. Differential diagnosis of intracranial masses. In J"urg Hodler, Rahel A. Kubik-Huch, and Johannes E. Roos, editors, *Diseases of the Brain, Head and Neck, Spine 2024–2027: Diagnostic Imaging*, pages 113–127. Springer, Cham, 2024. doi:10.1007/978-3-031-50675-8_8.

[23] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, 2017. doi:10.1038/sdata.2017.117.

[24] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. doi:10.1007/s10278-013-9622-7.

[25] Masoud Nickparvar. Brain tumor MRI dataset. Kaggle dataset, 2021. URL https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

[26] Burak Koçak. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: Focus on radiomics. *Diagnostic and Interventional Radiology*, 28(5):450–462, 2022. doi:10.5152/dir.2022.211297.

[27] Emma A. M. Stanley, Raissa Souza, Matthias Wilms, and Nils D. Forkert. Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. *eBioMedicine*, 111:105501, 2025. doi:10.1016/j.ebiom.2024.105501.

[28] Christian Tinauer, Maximilian Sackl, Rudolf Stollberger, Reinhold Schmidt, Stefan Ropele, and Christian Langkammer. Skull-stripping induces shortcut learning in MRI-based alzheimer's disease classification. *Insights into Imaging*, 16:283, 2025. doi:10.1186/s13244-025-02158-4.

[29] Nathan Drenkow, Mitchell Pavlak, Keith Harrigian, Ayah Zirikly, Adarsh Subbaswamy, Mohammad Mehdi Farhangi, Nicholas Petrick, and Mathias Unberath. Detecting dataset bias in medical AI using a generalized and modality agnostic auditing approach. *npj Digital Medicine*, 2026. doi:10.1038/s41746-026-02807-y.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pages 630–645, 2016. doi:10.1007/978-3-319-46493-0_38.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. doi:10.48550/arXiv.1409.1556.

[32] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017. doi:10.1109/CVPR.2017.243.

[33] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114, 2019. doi:10.48550/arXiv.1905.11946.

[34] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In *Computer Vision – ECCV 2020*, volume 12350 of *Lecture Notes in Computer Science*, pages 491–507, 2020. doi:10.1007/978-3-030-58558-7_29.

[35] Ross Wightman. Pytorch image models, 2019. URL https://github.com/huggingface/pytorch-image-models.

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 10347–10357, 2021. doi:10.48550/arXiv.2012.12877.

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. doi:10.1109/ICCV48922.2021.00986.

[38] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. doi:10.1109/ICCV48922.2021.00009.

[39] B. Sarada, K. Narasimha Reddy, R. Babu, and B. S. S. V. Ramesh Babu. Brain tumor classification using modified ResNet50V2 deep learning model. *International Journal of Computing and Digital Systems*, 16 (1):1–10, 2024. Preprint.

[40] Ahmad F. Al Bataineh, Kholoud M. O. Nahar, Hussein Khafajeh, Ghazi Samara, Rami Alazaidah, Aseel Nasayreh, Ahmad Bashkami, Haitham Gharaibeh, and Walaa Dawaghreh. Enhanced magnetic resonance imaging-based brain tumor classification with a hybrid Swin transformer and ResNet50V2 model. *Applied Sciences*, 14(22):10154, 2024. doi:10.3390/app142210154.

[41] Meryem Altin Karagoz, O. Ufuk Nalbantoglu, and Geoffrey C. Fox. Residual vision transformer (ResViT) based self-supervised learning model for brain tumour classification, 2024. URL https://doi.org/10.48550/arXiv.2411.12874.

[42] R. K. Yadav, M. Kumar, and A. Nandi. A scalable brain tumor diagnosis from large-scale MRI datasets using CNN–ViT and expert-attention fusions. *International Journal of Computational Intelligence Systems*, 18:254, 2025. doi:10.1007/s44196-025-00981-7.

[43] Seyed Amirali Zakavi, Parastoo Radnia, Mahdi Abdollahi Karizno, Armin Tafazolimoghadam, Hamidreza Amiri, Niloofar Deravi, and Mobina Fathi. Comparative evaluation of convolutional neural networks and Swin transformer for multiclass brain tumor MRI classification. *Discover Applied Sciences*, 2026. doi:10.1007/s42452-026-08767-y.

[44] Essam H. Houssein, Amr M. Gamal, Eman M. G. Younis, and Ebtsam Mohamed. Explainable artificial intelligence for brain tumor classification via fine-tuned transfer learning. *Discover Artificial Intelligence*, 6:306, 2026. doi:10.1007/s44163-026-00865-5.

[45] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[46] Tariq Shahzad, Sheikh Muhammad Saqib, Tehseen Mazhar, Weiwei Jiang, and Khmaies Ouahada. TumorSwin transformer model for brain tumor classification using MRI. *Array*, 30:100954, 2026. doi:10.1016/j.array.2026.100954.

[47] Tian Xia, Agisilaos Chartsias, and Sotirios A. Tsaftaris. Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. *Medical Image Analysis*, 64:101719, 2020. doi:10.1016/j.media.2020.101719.

[48] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLOS ONE*, 10(10):e0140381, 2015. doi:10.1371/journal.pone.0140381.

[49] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.02034.

[50] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

[51] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision*, pages 618–626, 2017. doi:10.1109/ICCV.2017.74.

[52] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018. doi:10.1109/WACV.2018.00097.

[53] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020. doi:10.18653/v1/2020.acl-main.385.

[54] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556, 2019. doi:10.18653/v1/N19-1357.

[55] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1810.03292.

[56] Nishanth Arun, Nathan Gaw, Praveer Singh, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021. doi:10.1148/ryai.2021200267.

[57] Ruitao Xie, Xiaoxi He, Limai Jiang, Mini Han Wang, Jingbang Chen, Rui Xiao, Bokai Yang, Ye Li, Jinling Tang, Yi Pan, and Yunpeng Cai. Bridging the interpretability gap for medical artificial intelligence models using class-association manifold learning. *Nature Biomedical Engineering*, 2026. doi:10.1038/s41551-026-01676-w.

[58] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

[59] Sivaramakrishnan Rajaraman, Poornachandran Ganesan, and Sameer Antani. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLOS ONE*, 17(1):e0262838, 2022. doi:10.1371/journal.pone.0262838.

[60] C. Penso, L. Frenkel, and J. Goldberger. Confidence calibration of a medical imaging classification system that is robust to label noise. *IEEE Transactions on Medical Imaging*, 43(6):2050–2060, 2024. doi:10.1109/TMI.2024.3353762.

[61] Shansong Wang, Mojtaba Safari, Qiang Li, Chih-Wei Chang, Richard L. J. Qiu, Justin Roper, David S. Yu, and Xiaofeng Yang. Vision foundation model for 3d magnetic resonance imaging segmentation, classification, and registration. *Medical Image Analysis*, 110:103992, 2026. doi:10.1016/j.media.2026.103992.

[62] Zhijian Yang, Noel DSouza, Istvan Megyeri, Xiaojian Xu, Amin Honarmandi Shandiz, Farzin Haddadpour, Krisztian Koos, Laszlo Rusko, Emanuele Valeriano, Bharadwaj Swaminathan, Lei Wu, Parminder Bhatia, Taha Kass-Hout, and Erhan Bas. Decipher-MR: A vision-language foundation model for 3d MRI representations. *npj Digital Medicine*, 9:419, 2026. doi:10.1038/s41746-026-02596-4.

[63] National Institute of Standards and Technology. Secure hash standard (SHS). Technical Report FIPS PUB 180-4, National Institute of Standards and Technology, 2015.

[64] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

[65] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

[66] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014. URL https://www.jmlr.org/papers/v15/srivastava14a.html.

[67] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[68] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

[69] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[70] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. doi:10.1109/CVPR.2016.308. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

[71] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1608.03983.

[72] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, 2013. URL https://proceedings.mlr.press/v28/pascanu13.html.

[73] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

[74] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) – Protein Structure*, 405(2):442–451, 1975. doi:10.1016/0005-2795(75)90109-9.

[75] Jan Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5–6):367–374, 2004. doi:10.1016/j.compbiolchem.2004.09.006.

[76] S. L. Taylor, J. A. Barakos, G. R. Harsh, and C. B. Wilson. Magnetic resonance imaging of tuberculum sellae meningiomas: Preventing preoperative misdiagnosis as pituitary macroadenoma. *Neurosurgery*, 31(4):621–627, 1992. doi:10.1227/00006123-199210000-00002.

[77] R. Kwancharoen, A. M. Blitz, F. Tavares, P. Caturegli, G. L. Gallia, and R. Salvatori. Clinical features of sellar and suprasellar meningiomas. *Pituitary*, 17(4):342–348, 2014. doi:10.1007/s11102-013-0507-z.

A Code-Derived Implementation

The notation follows Section 3; the order of gating, rotation, tokenisation, averaging and fusion follows the released Python definitions.

A.1 Fixed architectural settings

A.2 Hybrid A procedure

Hybrid A computes its CNN descriptor before applying PFD. The gated feature map is then rotated with its gate, converted into 49 tokens, weighted by the mean-normalised gate and supplied to the shared encoder. The CNN contribution to the fused classifier is therefore ungated.

A.3 Hybrid B procedure

Hybrid B applies PFD before its CNN descriptor is pooled. The same gate guides the image-patch pathway. The grid side is chosen once per batch from the unrotated gate, so every sample and every rotation has a common sequence length.

For batch index $b \in \{1, \dots, B\}$, flatten the mean-normalised 14×14 gate to $A_b = (A_{b,1}, \dots, A_{b,196})$ and define its token mean as $\overline{A}_b = \frac{1}{196} \sum_{i=1}^{196} A_{b,i}$. The code uses PyTorch's default sample standard deviation d_b and

Table 3: Architectural settings in the four released model definitions.

Setting	Hybrid A/ Ablation A	Hybrid B/ Ablation B
Input; final CNN map	$3 \times 224 \times 224$; $2048 \times 7 \times 7$	Same
Transformer input	Projected CNN feature map	Non-overlapping image patches
Initial token grid	$7 \times 7 = 49$	$14 \times 14 = 196$
Guided CNN descriptor	No in Hybrid A	Yes in Hybrid B
Rotation construction	Four aligned feature/gate rotations	Four aligned image/gate rotations
Embedding; depth; heads	142; 10; 10	142; 10; 10
Attention head width	14; internal width 140	15; internal width 150
RViT local operator	Depthwise 3×3 convolution	Same
Transformer MLP width	480	480
Branch and fusion width	256	256
Dropout	Transformer 0.1; fusion 0.5	Transformer 0.1; CNN/fusion 0.5
Observed parameters	26.68 million	26.58 million

then forms the batch statistic q:

$$d_b = \left[\frac{1}{196-1} \sum_{i=1}^{196} (A_{b,i} - \overline{A}_b)^2 \right]^{1/2}, \qquad q = \frac{1}{B} \sum_{b=1}^{B} d_b. \tag{19}$$

Equations (11) and (12) then determine the side and the guided token map. In Algorithms 1 and 2, $\mathrm{rot90}$ denotes a quarter-turn rotation, tokens spatial flattening, grid the inverse reshape, stack tensor stacking, interp positional-grid interpolation and $\mathrm{bilinear\_up}$ bilinear upsampling. The operator $\mathrm{PatchConv}_{16,16}$ is a 16×16 convolution with stride 16; RViT_{10} denotes ten residual RViT blocks followed by the encoder's final layer normalisation; and $\mathcal{A}$ denotes the collection of encoder attention matrices returned for explanation.

Figures 3 and 4 give the image-led forward paths; Algorithms 1 and 2 below record the same operations as executable pseudocode.

A.4 Ablation and compatibility details

The external embedding width is 142 and the head count is ten, following the adopted RViT configuration. Because 142 is not divisible by ten, the released Hybrid A attention uses $10 \times 14 = 140$ internal dimensions and Hybrid B uses $10 \times 15 = 150$; both project back to 142. Hybrid A also constructs its encoder with placeholder spatial metadata $7 // 16 = 0$. At run time, each block observes 49 tokens and reconstructs a 7×7 grid before its depthwise convolution. Both details affect exact reproduction.

B Reproducibility Details

Table 5: Training and evaluation settings used in the released experiments.

Component	Setting
Compute	One NVIDIA Tesla P100; two data-loader workers; pinned memory
Initialisation	Pretrained `resnetv2_50x1_bitm`; new layers use their module initialisers
Optimiser	AdamW; CNN learning rate 10^{-4}; transformer/fusion learning rate 5×10^{-4}
Regularisation	Weight decay 0.01 with stated exclusions; label smoothing 0.05; gradient norm at most 1.0
Schedule	Cosine annealing to 10^{-6}; maximum 100 epochs; early-stopping patience 10
Fine-tuning	CNN frozen for epochs 1–5, then unfrozen
Selection	Highest validation macro-F1; held-out test used after checkpoint selection
Batching	Batch 32; A family keeps the final incomplete batch; B family drops it
Augmentation	Rotation $\pm 15^\circ$; horizontal flip with probability 0.5; translation 5%; Gaussian noise with standard deviation 0.02 and probability 0.5
Normalisation	RGB tensor; per-channel mean 0.5 and standard deviation 0.5
Randomness	Seed 42; one run per model; Compute Unified Device Architecture (CUDA) Deep Neural Network library (cuDNN) benchmark enabled; strict determinism disabled
Precision	Standard precision; automatic mixed precision was not used
MC dropout	20 passes; dropout active; BN in evaluation mode

The repository contains the duplicate audit, fixed-split construction, four training configurations, model-specific explanation procedures, histories, confusion matrices, misclassification manifests, checkpoints and a reusable PFD–GSTE package. Raw images are obtained from the cited Kaggle source and are not redistributed. The one-run design means that the reported differences contain unknown optimisation variance.

C Complete Held-Out Results

The exact tables show that the common error is not a general failure to distinguish tumour from no tumour. Most remaining errors occur between tumour subtypes, and the direction

Algorithm 1 Hybrid A: PFD-A and static feature-token GSTE-A

Input: MRI batch x; rotations $\mathcal{K} = (0, 1, 2, 3)$.
Output: four-class logits $\widehat{y}$; optional gate and attention collection $\mathcal{A}$.

$F \leftarrow \mathcal{B}_\theta(x)$, with $F \in \mathbb{R}^{B\times 2048\times 7\times 7}$.
$z_{\text{cnn}} \leftarrow \text{ReLU}(W_c\,\text{GAP}(F) + b_c)$.
$M \leftarrow \sigma(W_m *_{1\times 1} F + b_m)$; $F^g \leftarrow F \odot M$.
Initialise the rotation-token list $\mathcal{T} \leftarrow [\,]$.
for (r, k) in enumerate$(\mathcal{K})$ **do**
  $F_r^g \leftarrow \text{rot90}(F^g, k)$; $M_r \leftarrow \text{rot90}(M, k)$.
  $E_r \leftarrow \text{tokens}(W_e *_{1\times 1} F_r^g + b_e) \in \mathbb{R}^{B\times 49\times 142}$.
  $a_r \leftarrow \text{tokens}(M_r)/(\text{mean}_i\,\text{tokens}(M_r)_i + 10^{-6})$.
  $T_r \leftarrow E_r \odot a_r + P_{7\times 7} + \rho_r$; append T_r to $\mathcal{T}$.
end for
$\overline{T} \leftarrow \text{mean}(\text{stack}(\mathcal{T}), \text{rotation axis})$.
$H, \mathcal{A} \leftarrow \text{RViT}_{10}(\overline{T})$.
$z_{\text{vit}} \leftarrow \text{ReLU}(W_v\,\text{mean}_i\,H_i + b_v)$.
$h \leftarrow \text{ReLU}(W_f[z_{\text{cnn}}; z_{\text{vit}}] + b_f)$.
$\widehat{y} \leftarrow W_o\,\text{Dropout}_{0.5}(h) + b_o$.
return $\widehat{y}$ and, when requested, up(M) and $\mathcal{A}$.

Algorithm 2 Hybrid B: PFD-B and dynamic patch-token GSTE-B

Input: MRI batch x; rotations $\mathcal{K} = (0, 1, 2, 3)$; base side $g = 14$.
Output: four-class logits $\widehat{y}$; optional gate, attention collection $\mathcal{A}$ and chosen side s.

$F \leftarrow \mathcal{B}_\theta(x)$.
$M \leftarrow \sigma(W_m *_{1\times 1} F + b_m)$; $F^g \leftarrow F \odot M$.
$M^\uparrow \leftarrow \text{bilinear\_up}(M, 224, 224)$.
$z_{\text{cnn}} \leftarrow \text{ReLU}(\text{BN}(W_c\,\text{Dropout}_{0.5}(\text{GAP}(F^g)) + b_c))$.
$A_0 \leftarrow \text{AAP}_{g\times g}(M^\uparrow)/(\text{mean}_{i,j}\,\text{AAP}_{g\times g}(M^\uparrow)_{i,j} + 10^{-6})$.
$q \leftarrow \text{mean}_b(\text{std}_{i,j}(A_{0,b}; \text{sample}))$.
Choose $s \in [7, 14]$ from q using Equation (11).
Initialise the rotation-token list $\mathcal{T} \leftarrow [\,]$.
for k in $\mathcal{K}$ **do**
  $x_k \leftarrow \text{rot90}(x, k)$; $E_k \leftarrow \text{PatchConv}_{16,16}(x_k) \in \mathbb{R}^{B\times 196\times 142}$.
  $M_k \leftarrow \text{rot90}(M^\uparrow, k)$.
  $A_k \leftarrow \text{AAP}_{g\times g}(M_k)/(\text{mean}_{i,j}\,\text{AAP}_{g\times g}(M_k)_{i,j} + 10^{-6})$.
  $X_k \leftarrow \text{grid}(E_k) \odot A_k$.
  if $s < g$ **then** $X_k \leftarrow \text{AAP}_{s\times s}(X_k)/(\text{AAP}_{s\times s}(A_k) + 10^{-6})$.
  $T_k \leftarrow \text{tokens}(X_k) + \text{interp}(P_{14\times 14}, s, s) + \rho_k$; append T_k to $\mathcal{T}$.
end for
$\overline{T} \leftarrow \text{mean}(\text{stack}(\mathcal{T}), \text{rotation axis})$.
$H, \mathcal{A} \leftarrow \text{RViT}_{10}(\overline{T})$.
$z_{\text{vit}} \leftarrow \text{ReLU}(W_v\,\text{mean}_i\,H_i + b_v)$.
$h \leftarrow \text{ReLU}(W_f[z_{\text{cnn}}; z_{\text{vit}}] + b_f)$.
$\widehat{y} \leftarrow W_o\,\text{Dropout}_{0.5}(h) + b_o$.
return $\widehat{y}$ and, when requested, $M^\uparrow$, $\mathcal{A}$ and s.

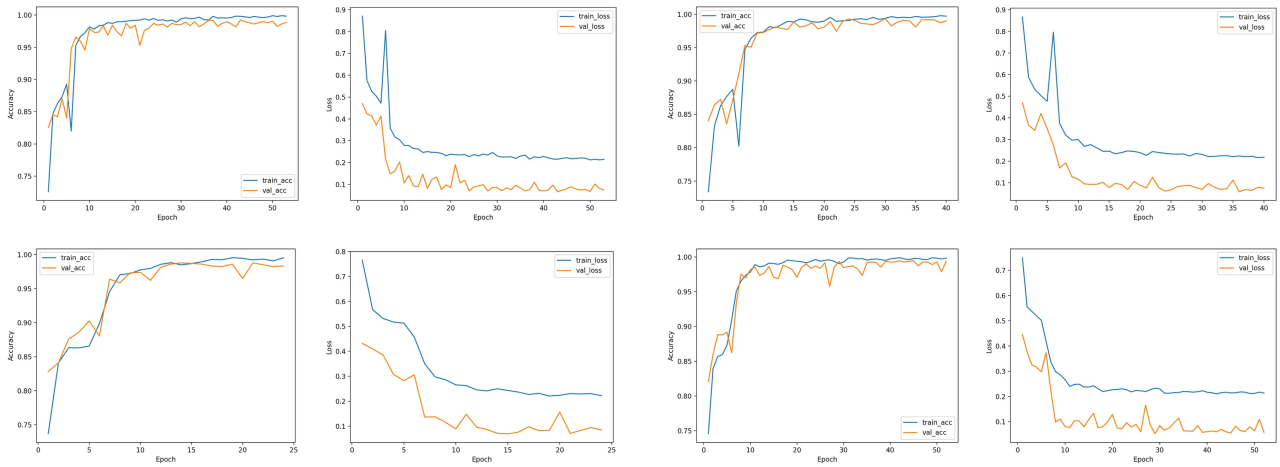
Table 4: The operations removed by each ablation and those retained.

Control	Removed	Retained
Ablation A	PFD gate; feature-token weights	ResNet50V2 descriptor; four rotations of the ungated 7×7 feature map; 49-token encoder; late fusion
Ablation B	PFD gate; gated CNN descriptor; patch-token weights; grid reduction	ResNet50V2 descriptor; four raw-image rotations; full 196-token encoder; late fusion

differs by model. The highest no-tumour false-positive count is seven in Hybrid B; the highest single subtype confusion is seven meningioma images predicted as pituitary by Ablation A.

D Qualitative Explanation and Uncertainty Audit

No tumour masks, overlap scores, pointing-game tests, causal deletion tests, randomisation tests or blinded clinical readers were available. The observations of 0/3, 3/3, 1/3 and 1/3 tumour-centred Grad-CAM++ cases are therefore



(a) Guided-model histories.

(b) Ablation histories.

Figure 10: Released training histories. Checkpoints are selected by validation macro-F1: Hybrid B at epoch 14, Hybrid A at epoch 43, Ablation A at epoch 30 and Ablation B at epoch 42.

Table 6: Per-class precision (P), recall (R), F1 and one-versus-rest specificity (S) on the 1,284-image test set.

Model	Class	P	R	F1	S
Hybrid A	Glioma	0.9801	0.9900	0.9850	0.9939
	Meningioma	0.9837	0.9901	0.9869	0.9949
	Pituitary	0.9900	0.9900	0.9900	0.9970
	No tumour	0.9947	0.9816	0.9881	0.9978
Hybrid B	Glioma	0.9705	0.9900	0.9801	0.9909
	Meningioma	0.9710	0.9901	0.9805	0.9908
	Pituitary	0.9966	0.9800	0.9882	0.9990
	No tumour	1.0000	0.9816	0.9907	1.0000
Ablation A	Glioma	0.9835	0.9967	0.9900	0.9949
	Meningioma	0.9867	0.9770	0.9818	0.9959
	Pituitary	0.9770	0.9933	0.9851	0.9929
	No tumour	1.0000	0.9843	0.9921	1.0000
Ablation B	Glioma	0.9933	0.9866	0.9899	0.9980
	Meningioma	0.9806	0.9967	0.9886	0.9939
	Pituitary	0.9933	0.9933	0.9933	0.9980
	No tumour	1.0000	0.9921	0.9960	1.0000

Table 7: Exact confusion counts. Rows are true classes and columns are predicted classes: glioma (G), meningioma (M), pituitary (P) and no tumour (N).

Hybrid A

True	G	M	P	N
G	296	1	2	0
M	0	301	1	2
P	1	2	297	0
N	5	2	0	374

Hybrid B

True	G	M	P	N
G	296	3	0	0
M	2	301	1	0
P	1	5	294	0
N	6	1	0	374

Ablation A

True	G	M	P	N
G	298	1	0	0
M	0	297	7	0
P	1	1	298	0
N	4	2	0	375

Ablation B

True	G	M	P	N
G	295	3	1	0
M	0	303	1	0
P	0	2	298	0
N	2	1	0	378

descriptive. A plausible heatmap does not establish that the highlighted pixels caused the final prediction.

Table 8: Visual interpretation of the three fixed cases. GC denotes Grad-CAM++; AR denotes Attention Rollout. These are qualitative judgements, not localisation scores.

Case	Model	GC	AR
Correct meningioma	Hybrid A	Wrong	Partly correct
	Hybrid B	Tumour-centred	Wrong
	Ablation A	Wrong	Partly correct
	Ablation B	Tumour-centred	Wrong
Challenging glioma	Hybrid A	Wrong	Wrong
	Hybrid B	Tumour-centred	Partly wrong
	Ablation A	Tumour-centred	Partly wrong
	Ablation B	Partly wrong	Partly wrong
Pituitary → meningioma	Hybrid A	Wrong	Wrong
	Hybrid B	Tumour-centred	Partly correct
	Ablation A	Wrong	Partly correct
	Ablation B	Wrong	Partly correct

Scope. The released evidence supports four-class results on one fixed, exact-duplicate-audited Kaggle split; a code-defined comparison of PFD–GSTE against two family-specific ablations; and a qualitative explanation audit of three test cases. It does not support patient-independent generalisation, quantitative tumour localisation, calibrated uncertainty or clinical use. Those conclusions require patient identifiers, external cohorts, repeated runs, tumour masks, formal localisation tests and probability-level calibration analysis.

Table 9: MC dropout mean confidence on the shared error. Every model predicts meningioma; the true class is pituitary.

Model	Mean confidence	Passes	Prediction	True class
Hybrid A	0.9692	20	Meningioma	Pituitary
Hybrid B	0.9869	20	Meningioma	Pituitary
Ablation A	0.9592	20	Meningioma	Pituitary
Ablation B	0.9632	20	Meningioma	Pituitary