UNIVERSITY OF HERTFORDSHIRE
Department of Computer Science

Modular BSc Honours in Computer Science
(Artificial Intelligence)

6COM2017 – Artificial Intelligence Project

Final Report

26 March 2026

Mitigating Shortcut Learning in Brain Tumour MRI Classification

Author: Riya Basak

Supervised by: Dr Kheng Lee Koay

Abstract

This project investigated whether *shortcut learning* could be reduced in brain tumour *Magnetic Resonance Imaging* (MRI) classification. Here, shortcut learning means correct predictions made for the wrong visual reasons, such as background rather than tumour-related cues. Two guidance modules, *Pathology-Focused Disentanglement* (PFD) and *Guided Semantic Token Evolution* (GSTE), were added to hybrid *Convolutional Neural Network* (CNN) and *Rotation-Invariant Vision Transformer* (RViT) models. Four models were evaluated: *Hybrid A, Hybrid B, Ablation A*, and *Ablation B.*

Hybrid A applied *weaker guidance to 7×7 CNN-derived tokens*, while Hybrid B applied *stronger guidance to raw-image patch tokens and the CNN descriptor*; Ablation A and Ablation B kept the same family structures but removed PFD and GSTE. A reproducible four-class benchmark was prepared using offline preprocessing, duplicate removal, and fixed train, validation, and test splits. Evaluation used classification metrics, confusion matrices, *Gradient-weighted Class Activation Mapping Plus Plus* (Grad-CAM++), *Attention Rollout,* and *Monte Carlo dropout uncertainty*.

All models achieved high test accuracy (*0.9852-0.9922*). *Ablation B achieved the best performance*, suggesting that the more direct RViT-based B-family design was stronger for classification, whereas Hybrid B produced the most tumour-centred Grad-CAM++ results, indicating that stronger guidance improved evidence focus more clearly than final classification performance.

Acknowledgements

I would like to thank my supervisor, Dr Kheng Lee Koay, for his guidance, patience, and clear feedback throughout this project. His feedback helped me recognise gaps in my understanding, think more deeply about the work, and make better progress at each stage. Even when I was not able to explain my ideas clearly, he was patient in understanding what I meant and helped me find a clearer direction. I am also grateful to my module leaders and tutors for their teaching and support since
Year 1, which has contributed greatly to my development. I would also like to thank my family for always being there for me, listening to me, and supporting me throughout my studies. Finally, I thank the SASH team for their helpful feedback in improving the report and making its structure clearer.

Table of Contents

Abstract ii
Acknowledgements iii
Table of Contents iv
Abbreviations vi
Glossary vii
1. Introduction 1
1.1 Project idea and motivation 1
1.2 Project Aims, Objectives, and Targets 1
1.3 Report structure 2
2. Literature Review 3
2.1 Introduction to the Clinical Problem: Brain Tumours, Classification and Diagnostics 3
2.1.1 Clinical and imaging context 4
2.2 Dataset landscape and justification for selecting a feasible benchmark 4
2.3 Deep learning foundations for classification: CNNs and Transformers 5
2.4 CNN backbones for tumour MRI classification 5
2.4.1 Why ResNet50V2 (and not VGG/DenseNet/EfficientNet) 5
2.5 Vision Transformers and their variants 5
2.5.1 Why RViT (and not DeiT/Swin/CvT) 5
2.6 Hybrid CNN-Transformer systems: motivation, patterns, and risks 7
2.7 Prior work motivating pathology-centred focus and token efficiency 7
2.8 Explainability and uncertainty for human-centred evaluation 8
2.9 Synthesis: gap and project justification 9
3. Project Method 10
3.1 Legal, ethical and professional considerations 10
3.1.1 Dataset source and provenance 10
3.2 Dataset description and label structure 10
3.2.1 Dataset curation and preprocessing pipeline 11
3.2.2 Data splitting and test-set protection 12
3.2.3 Dataset interface and online transformations 13
3.3 Baselines, alternatives considered, and rationale 13
3.3.1 Options considered but not addressed 13
3.4 System architecture: hybrid CNN-Transformer 13
3.4.1 Hybrid A 14
3.4.2 Hybrid B 14
3.4.3 Architectural comparison 14
3.4.4 Proposed contributions: PFD and GSTE 17
3.5 RViT branch implementation and relation to Krishnan et al. (2024) 17
3.5.1 Fusion of ResNet50V2 and RViT branches 18
3.6 Explainability integration 18
3.6.1 Uncertainty estimation 18
3.7 Ablation study and supporting web interface 19

4. Training 20
4.1 Training objective and comparisons 20
4.1.1 Training-time input processing: augmentation and normalisation 20
4.1.2 Optimisation strategy 20
4.1.3 Stability controls and implementation details 20
4.1.4 Model selection and checkpoint saving 21
4.1.5 Evaluation metrics and reporting 21
4.2 Training dynamics and convergence 21
5. Results 24
5.1 Overall held-out test performance 24
5.1.1 Per-class performance 24
5.1.2 Confusion matrices 25
5.1.3 Error analysis with misclassification counts 26
5.1.4 Compact ablation difference summary 26
5.2 Explainability (XAI) results: Grad-CAM++ and Attention Rollout 26
5.2.1 Comparison setup 26
5.2.2 Correct-case comparison 27
5.2.3 Challenging-case comparison 27
5.2.4 Visual and confidence summary 29
5.3 Uncertainty and confidence behaviour 29
6. Discussion and Evaluation 30
6.1 Main outcome 30
6.2 What the explainability results mean 30
6.3 Errors and uncertainty 30
6.4 Comparison with existing benchmark studies and research gap 31
6.5 Limitations and overall evaluation 32
7. Conclusion 32
8. Future Work 34
9. Project Management Review 35
10. References 38
11. Bibliography 44
12. Appendices 45
A1. Training and model hyperparameters for Hybrid models and their ablations 45
A2. External software libraries, frameworks and implementation tools 46
A3. Source code access and repository guide 46
A4. Reproducibility and setup guide 47

Abbreviations

Abbreviation	Full form
AI	Artificial Intelligence
AdamW	Adaptive Moment Estimation with decoupled weight decay
AMP	Automatic Mixed Precision
BraTS	Brain *Tumor* Segmentation
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
CLS token	Classification token
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture
EXIF	Exchangeable Image File Format
F1	Harmonic mean of precision and recall
FLAIR	Fluid-Attenuated Inversion Recovery
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
GSTE	Guided Semantic Token Evolution
ICO	Information Commissioner's Office
MCC	Matthews Correlation Coefficient
MC Dropout	Monte Carlo dropout
MF1	Macro-F1
MRI	Magnetic Resonance Imaging
OOD	Out-of-distribution
PFD	Pathology-Focused Disentanglement
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
RViT	Rotation-Invariant Vision Transformer
SHA1	Secure Hash Algorithm 1
T1	T1-weighted MRI
T1ce	T1-weighted post-contrast MRI
T2	T2-weighted MRI
TCIA	The Cancer Imaging Archive
ViT	Vision Transformer
WF1	Weighted-F1
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

Glossary

Term	Meaning
ablation	A comparison in which a model component is removed to test its contribution.
acquisition bias	Patterns linked to scanner type, hospital workflow, or image settings rather than disease itself.
attention rollout	A transformer explanation method that estimates which tokens influenced the decision by tracing attention across layers.
augmentation	Random image changes applied during training to improve generalisation.
benchmark	A dataset and evaluation setting used for method comparison.
calibration	The extent to which model confidence matches actual correctness.
checkpoint	A saved model state used for later evaluation or inference.
confusion matrix	A table showing how often each true class is predicted as each possible class.
cross-entropy loss	A standard classification loss that penalises incorrect predictions.
data leakage	When information from the test set influences model development.
early stopping	Stopping training when validation performance no longer improves.
feature map	The spatial output produced by a convolutional neural network.
false negative	A tumour case incorrectly predicted as no tumour.
false positive	A no tumour case incorrectly predicted as tumour.
fixed split	Reuse of the same train, validation, and test sets across experiments.
gating	Applying weights so some regions influence the model more than others.
Gantt chart	A project management chart showing tasks across time.
Grad-CAM++	A CNN explanation method that highlights image regions contributing to a prediction.
gradient clipping	Limiting gradient size to improve training stability.
held-out test set	Data kept separate until final evaluation.
inductive bias	A built-in tendency that makes some patterns easier for a model to learn.
label smoothing	A training method that softens hard class labels to reduce overconfidence.
late fusion	Combining branch outputs near the end of a model.
localisation	The extent to which an explanation map focuses on the relevant region.
patch	A small image region used as an input unit for a transformer.
precision	Of the cases predicted as a class, the proportion that are correct.
preprocessing	Steps applied before training, such as cropping, resizing, RGB conversion, and duplicate checking.
recall / sensitivity	Of the cases that truly belong to a class, the proportion correctly detected.
shortcut learning	Reliance on easy but non-causal cues rather than genuine disease evidence.
specificity	The ability to correctly identify cases that do not belong to a class.
stratified split	A data split designed to preserve class proportions.
token	A unit of representation used by a transformer.
token-grid shrinking	Reducing the number of transformer tokens by pooling them into a smaller grid.
transfer learning	Adapting a model pretrained on another dataset to a new task.
uncertainty	The degree to which the model appears unsure about a prediction.
validation set	Data used during development for model selection, but not for final test reporting.
warmup	An initial stage of training used to stabilise optimisation before full fine-tuning.
weighted pooling	Combining features or tokens using learned or guidance-based weights.

1. Introduction

1.1 Project idea and motivation

Brain tumours are abnormal cell growths in the brain or nearby central nervous system structures (the parts of the body that include the brain and spinal cord) (Louis et al., 2021). Accurate brain tumour classification is important because errors or delays can affect treatment planning, follow-up decisions, and patient outcomes. Magnetic Resonance Imaging (MRI) is central to brain tumour assessment, but classification remains difficult because tumour appearance can overlap across subtypes and can vary with slice position, orientation, and imaging conditions (Louis et al., 2021; Villanueva-Meyer, Mabray and Cha, 2017).
This makes brain tumour MRI classification a relevant problem for Artificial Intelligence (AI)-based decision support.

Recent studies show that Convolutional Neural Networks (CNNs), transformers, and hybrid models can achieve strong results on the Nickparvar (2021) Kaggle Brain *Tumor* MRI dataset. On four-class brain tumour classification, reported results include 96.33% for a modified ResNet50V2 model, 96.8% for a ResNet50V2-Swin hybrid, and 98.47% for a Residual Vision Transformer (ResViT)-based hybrid CNN-transformer model (Sarada et al., 2025; Al Bataineh et al., 2024; Karagoz et al., 2024).
By contrast, Rotation-Invariant Vision Transformer (RViT) was reported at 98.6% on a binary glioma-against-no tumour task using the same dataset (Krishnan et al., 2024).

This matters because strong benchmark accuracy alone does not guarantee that a model is using the tumour itself. A model may instead rely on shortcut cues, such as background patterns, or other dataset-specific regularities, while still appearing accurate on the benchmark (Geirhos et al., 2020; Roberts et al., 2021). Prior studies identified for this project had not yet reported a ResNet50V2-RViT fusion in this classification setting, nor had they directly tested whether added guidance could reduce this risk.

This project therefore investigates a hybrid inspired by ResNet50V2 and RViT, using the same Nickparvar (2021) dataset, and introduces two guidance modules: Pathology-Focused Disentanglement (PFD) and Guided Semantic Token Evolution (GSTE), in weaker and stronger variants with their ablations, to test whether stronger tumour-focused guidance can help mitigate shortcut learning while maintaining reliable classification performance.

Explainable Artificial Intelligence (XAI) and uncertainty analysis were included to examine not only whether the models were correct, but also whether they focused on tumour-related regions and whether their predictions were made with stable confidence. Gradient-weighted Class Activation Mapping Plus Plus (Grad-CAM++) was used for the CNN branch, Attention Rollout for the transformer branch, and Monte Carlo (MC) Dropout for uncertainty analysis.

The research question was: to what extent can pathology-guided hybrid CNN-transformer models for four-class brain tumour MRI classification maintain strong test performance while producing more tumour-focused explanations and more informative uncertainty behaviour than similar versions without the guidance modules?

1.2 Project Aims, Objectives, and Targets

1.2.1 Aim

To implement and evaluate hybrid CNN-transformer models for four-class brain tumour MRI classification, and to test whether guidance modules can improve tumour-focused evidence and prediction uncertainty behaviour as an experimental approach to reducing shortcut learning.

1.2.2 Objectives

- To prepare a reproducible four-class MRI benchmark from the Kaggle Brain *Tumor* MRI dataset using offline preprocessing, duplicate checking, and fixed train, validation, and test splits (Nickparvar, 2021).
- To implement two hybrid variants (Hybrid A and Hybrid B) and two ablation models in which the guidance modules are removed.
- To train all four variants under the same optimisation and model-selection procedure.
- To evaluate test performance using standard classification metrics, confusion matrices, and misclassification counts.
- To compare model focus using Grad-CAM++ for the CNN branch and Attention Rollout for the transformer branch.
- To analyse prediction confidence using MC Dropout to distinguish uncertain errors from wrong-but-confident errors (Gal and Ghahramani, 2016).

1.2.3 Targets

- To train and evaluate all four variants successfully under one shared train, validation, and test split.
- To achieve strong test performance, with at least one variant reaching a test macro-F1 (unweighted average F1 across classes) of 0.98 or higher.
- To produce ablation results so that the effect of PFD-GSTE can be examined fairly.
- To generate explainability outputs on a fixed set of correct and incorrect test examples.
- To report uncertainty summaries alongside selected success and failure cases.

1.3 Report structure

This report is organised into nine main chapters:

- Chapter 1 (Introduction): outlines the project background, motivation, research question, aims, objectives, and targets.
- Chapter 2 (Literature Review): reviews the main clinical and technical studies related to brain-tumour MRI classification, including datasets, CNNs, transformers, hybrid models, explainability, uncertainty, and the research gap addressed in this project.
- Chapter 3 (Project Method): explains the dataset, preprocessing, duplicate checking, test-set protection, hybrid architectures, proposed PFD and GSTE modules, and the overall experimental design.
- Chapter 4 (Training): describes the shared training procedure, data augmentation, optimisation settings, early stopping, and model selection.
- Chapter 5 (Results): presents the held-out test results, per-class performance, confusion matrices, error analysis, explainability examples, and uncertainty outputs.
- Chapter 6 (Discussion and Evaluation): interprets the findings in relation to the project aims, including performance, evidence quality, and limitations.
- Chapter 7 (Conclusion): summarises the main findings and answers the research question.
- Chapter 8 (Future Work): suggests the main directions for further improvement and study.
- Chapter 9 (Project Management Review): reflects on project planning, progress, challenges, and lessons learned.

2. Literature Review

This chapter reviews the literature relevant to brain tumour MRI classification, including datasets, CNN and transformer approaches, hybrid design, shortcut learning, explainability, and uncertainty. Its purpose is to identify the main strengths and limits of existing work and to show how these lead to the present study.

2.1 Introduction to the Clinical Problem: Brain Tumours, Classification and Diagnostics

Brain tumours are clinically important because both non-cancerous (benign) and cancerous (malignant) tumours can be dangerous. Even benign tumours can harm vision, thinking or hormone control if they press on important parts of the brain within the limited space of the skull
(the hard bone around the brain), while malignant tumours are usually more invasive, faster growing and linked to poorer survival (Louis et al., 2016; Louis et al., 2021). Figure 2.1(b) shows example MRI appearances of benign and malignant brain tumours. Under the World Health Organization (WHO) central nervous system framework, grade 1 tumours are generally less aggressive, whereas grade 3 and grade 4 tumours, including glioblastoma, are usually more serious and linked to worse outcomes; some tumours that begin as benign, especially meningiomas, can return and only rarely become malignant, while pituitary adenomas are usually slow growing and only rarely become malignant.
In this report, *pituitary* is used as a short label for pituitary adenoma (Louis et al., 2021; Maggio et al., 2021; Ho et al., 2023). Figure 2.2 shows representative MRI examples of glioma, meningioma, pituitary adenoma and no tumour.

Brain tumours affect people across the lifespan differently: they are the most common solid tumours in children, adult diffuse gliomas are a major tumour group in adults, and meningiomas become more common in older adults (Pollack et al., 2019; Weller et al., 2021; Amoo et al., 2023). Common symptoms include headache, seizures, weakness, changes in thinking, nausea and vomiting, but these signs are often non-specific and can overlap across tumour types (Comelli et al., 2017; Weller et al., 2021). MRI is central to diagnosis because it shows brain soft tissue clearly, but tumour subtypes can still look similar, and scans can vary by position, orientation and acquisition setting (Gholipour et al., 2007). Figure 2.1(a) shows how MRI scans can differ in brain view, position, orientation and field of view.

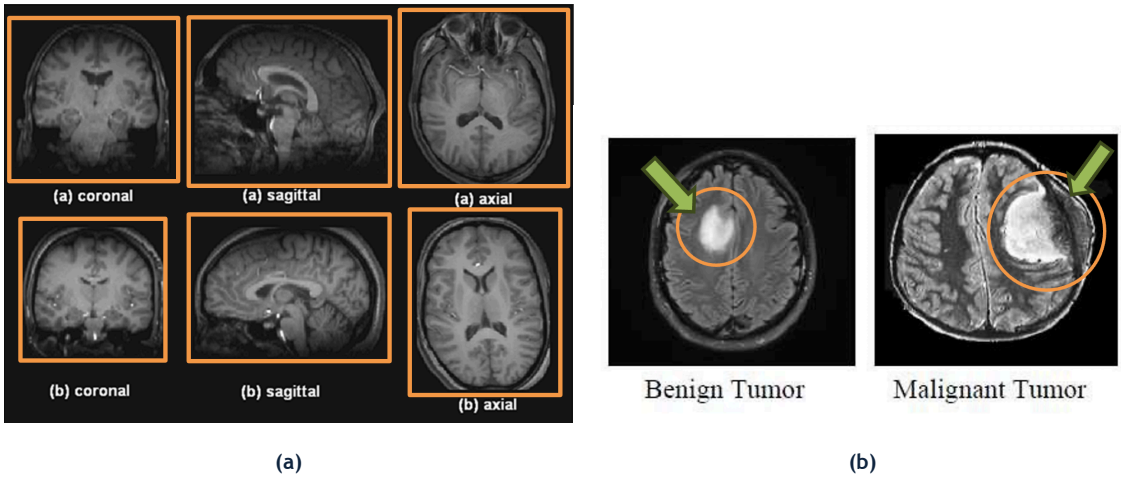


Figure 2.1: Left (a): MRI scans of two subjects with differing position, orientation and field of view, illustrating the need for spatial normalisation. Adapted from Gholipour et al. (2007, fig. 7, p. 438); the 3D brain views are outlined to make the differences in view and positioning easier to see. **Right (b):** Examples of benign and malignant brain tumours on MRI, with tumour regions highlighted to indicate typical locations. Adapted from Khairandish, Gupta and Sharma (2020, fig. 3, p. 6869); the tumour regions are circled and arrows indicate their locations.

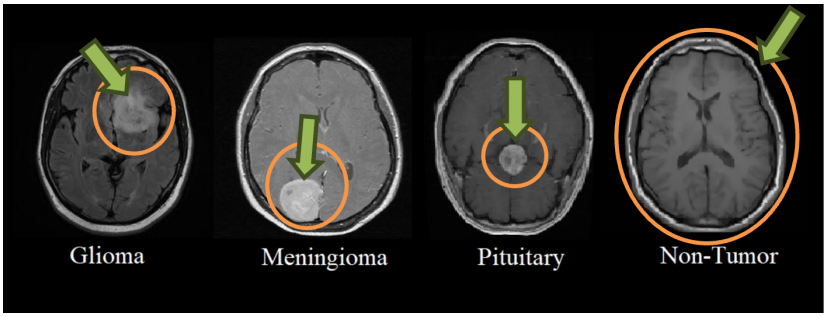


Figure 2.2: Representative MRI examples of major brain tumour types, including glioma, meningioma, pituitary and no tumour. Adapted from Yazdan et al. (2022, fig. 1, p. 1911); the regions of interest are circled and arrows indicate the tumour and no tumour locations.

2.1.1 Clinical and imaging context

MRI sequence choice affects how tumours appear and therefore what a model can learn. T1-weighted images show anatomy clearly, contrast-enhanced T1-weighted images (T1ce) make some tumour regions stand out more clearly, T2-weighted images highlight fluid-related signal, and FLAIR (Fluid-Attenuated Inversion Recovery) suppresses cerebrospinal fluid while preserving oedema visibility (swelling around the tumour) (Villanueva-Meyer, Mabray and Cha, 2017; Bakas et al., 2017).

Figure 2.3 illustrates the main MRI sequences used in brain tumour imaging and how tumour appearance can differ across them. This matters because a model trained on one imaging style may learn scanner or sequence-specific cues instead of pathology (abnormal tumour area), creating weaker real-world generalisation.

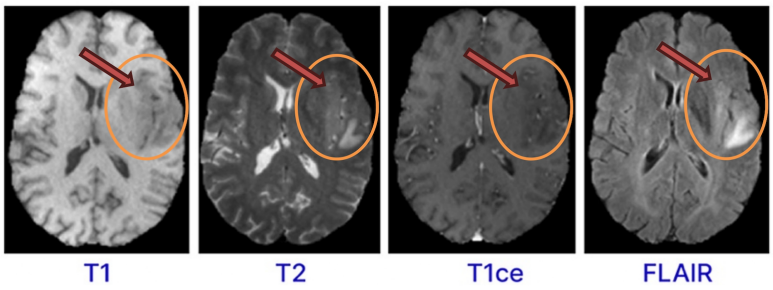


Figure 2.3: MRI sequences commonly used in brain tumour imaging (T1, T1ce, T2 and FLAIR), with tumour region indicated. Adapted from Farhan, Khalid and Manzoor (2025, fig. 1, p. 3); the tumour region is highlighted in each sequence to show how tumour appearance changes across MRI types.

2.2 Dataset landscape and justification for selecting a feasible benchmark

Brain-tumour MRI datasets differ in both modality and task design. Brain *Tumor* Segmentation (BraTS) datasets provide multi-modal 3D MRI with segmentation labels and greater clinical richness, but they require heavier computation and more complex evaluation than a 2D classification study (Bakas et al., 2017). The Cancer Imaging Archive (TCIA) collections offer more heterogeneous real-world imaging, but often require careful adjustment and do not fit neatly into a simple four-class 2D classification task (Clark et al., 2013).

For this project, the Kaggle Brain *Tumor* MRI benchmark was selected because it supports reproducible four-class classification experiments and model comparison under limited compute. It is commonly described as an aggregated 2D dataset of 7,023 images across glioma, meningioma, pituitary tumour, and no tumour classes, with tumour images largely derived from T1ce MRI sources such as Figshare (Cheng et al., 2015; Nickparvar, 2021). However, because it is aggregated, it carries a higher risk of data leakage (information from the held-out test set reaches model development), duplicate overlap, and dataset bias caused by differences in scanner type, hospital source, or image acquisition (Roberts et al., 2021; Koçak, 2022).

2.3 Deep learning foundations for classification: CNNs and Transformers

CNNs remain widely used in medical image classification because convolution gives a strong locality bias (a built-in focus on nearby image patterns), which often makes CNNs more data-efficient on moderate-sized datasets (LeCun, Bengio and Hinton, 2015). Transfer learning is also common, since ImageNet-pretrained CNNs can stabilise training when labelled medical data are limited (Shin et al., 2016; Tajbakhsh et al., 2016).

Transformers model wider spatial relationships through self-attention rather than fixed local filters (Vaswani et al., 2017). Vision Transformer (ViT) applies this idea to images by processing patch tokens as a sequence (Dosovitskiy et al., 2021). This improves global context modelling, but usually increases computational cost and dependence on stronger regularisation. The literature therefore suggests a clear trade-off: CNNs are stronger for local feature extraction, while transformers are better at wider context modelling. This motivates hybrid design for the present study.

2.4 CNN backbones for tumour MRI classification

Several CNN families have been used in medical image classification, including Visual Geometry Group (VGG), Inception, DenseNet, EfficientNet, and ResNet (Simonyan and Zisserman, 2015; Szegedy et al., 2015; Huang et al., 2017; He et al., 2016; Tan and Le, 2019). In brain-tumour MRI studies, however, backbone (the main base network that extracts image features) choice cannot be judged from overall accuracy alone, because reported performance also depends on dataset composition, split policy, augmentation, and fine-tuning strategy (Koçak, 2022). For this study, the CNN branch needed a stable and practical feature extractor that could support reliable local feature learning within the hybrid model and across ablations. A suitable backbone was therefore one that was widely used, stable, and feasible within the project's compute and time limits.

2.4.1 Why ResNet50V2 (and not VGG/DenseNet/EfficientNet)

ResNet50V2 was selected because this study was partly inspired by Sarada et al. (2025), who used a modified ResNet50V2 for brain-tumour MRI classification. In the present study, ResNet50V2 was used as the CNN backbone because it provides strong local feature extraction, stable transfer learning, and a practical base for repeated hybrid and ablation experiments under limited compute. VGG is heavier, DenseNet increases memory cost, and EfficientNet introduces extra scaling choices that were not central to this investigation (Simonyan and Zisserman, 2015; Huang et al., 2017; Tan and Le, 2019). ResNet50V2 was therefore chosen as a stable and suitable CNN branch so that the main focus could remain on testing the hybrid design and the proposed PFD-GSTE guidance.

2.5 Vision Transformers and their variants

Vision Transformers (ViTs) model global image relationships through self-attention over patch tokens, which can help capture wider spatial structure than CNNs alone (Dosovitskiy et al., 2021; Vaswani et al., 2017). However, this strength comes with practical limits: ViTs are often more data-hungry, more sensitive to training setup, and more computationally expensive as token count increases. In moderate-sized MRI datasets, this makes it difficult to tell whether gains come from genuine pathology learning or from dataset-specific shortcuts.

Several variants attempt to address these weaknesses. Data-efficient Image Transformers (DeiT) improves data efficiency, Swin Transformer (Swin) introduces windowed attention for better efficiency and locality, and Convolutional Vision Transformer (CvT) adds convolutional structure to strengthen inductive bias (a built-in tendency to learn certain patterns more easily) (Touvron et al., 2021; Liu et al., 2021; Wu et al., 2021). For brain-tumour MRI, however, another issue is especially relevant:
scan orientation can vary across acquisition and preprocessing. Krishnan et al. (2024) address this directly with RViT, which integrates information across rotated views.

2.5.1 Why RViT (and not DeiT/Swin/CvT)

RViT was selected because it directly addresses the transformer-side issue most relevant to this project: MRI scans may appear in different orientations, and the model needs to cope with that reliably

(Krishnan et al., 2024). DeiT, Swin, and CvT improve efficiency or general image modelling, but handling rotation is not their main design goal (Touvron et al., 2021; Liu et al., 2021; Wu et al., 2021). Swin and CvT also bring extra design choices, which would make the comparison less focused and harder to interpret.
RViT was therefore chosen not because it is assumed to be the best transformer in every setting, but because it is more directly relevant to this project's problem (Koçak, 2022; Krishnan et al., 2024). In this project, RViT is used as the transformer-side design principle so that rotation-aware global information can be added to the hybrid model and judged through fair comparison, rather than overall test accuracy alone.

Figure 2.4(a) by Sarada et al. (2025) shows a CNN-based design in which the MRI passes through convolution and residual blocks, with batch normalisation, max-pooling, and dropout supporting training stability and classification. Figure 2.4(b) by Krishnan et al. (2024) shows a transformer-based design in which the image is split into patches, rotated copies are created, patch embeddings from these views are combined, and self-attention is used to learn relationships between patches before classification. ResNet50V2 focuses on local texture and lesion-shape features (the shape of the abnormal area in the scan), whereas RViT focuses on wider patch-to-patch relationships with added rotation awareness. This is why both papers are relevant to this project: one motivates the CNN branch and the other motivates the transformer branch.

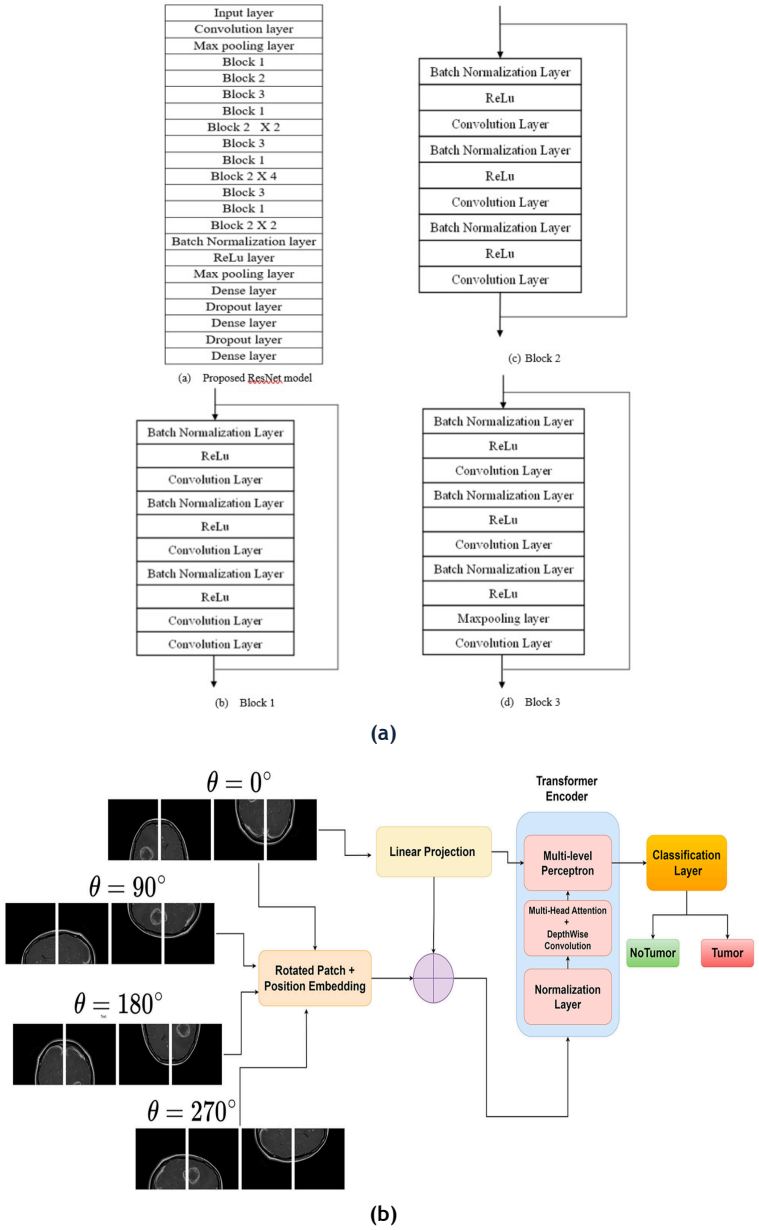


Figure 2.4: Upper (a): Modified ResNet50V2 brain tumour classification architecture. Reproduced from Sarada et al. (2025, fig. 2, p. 5). **Lower (b):** Rotation-Invariant Vision Transformer (RViT) architecture. Reproduced from Krishnan et al. (2024, fig. 1, p. 4).

2.6 Hybrid CNN-Transformer systems: motivation, patterns, and risks

Hybrid CNN-transformer models are relevant to brain-tumour MRI classification, and the comparative rationale is summarised in Table 2.1. Recent studies support this direction: Al Bataineh et al. combine ResNet50V2 with a Swin Transformer, and Karagoz et al. use a ResViT-based hybrid under limited-data conditions (Al Bataineh et al., 2024; Karagoz et al., 2024). However, these studies mainly focus on predictive performance and do not examine reliability more closely through ablation, branch-specific explanation, and uncertainty analysis. Many hybrid studies also report final accuracy without clearly showing what each branch contributes to the decision. The present project therefore treats hybrid design as a testable choice rather than an assumption.

Table 2.1: Technological justification (CNN, ViT and the hybrid)

Feature	CNN family (ResNet50V2 structure)	ViT family (ViT / RViT structure)	Why the hybrid is justified here
Core operation	Convolution with residual connections learns local spatial patterns efficiently (He et al., 2016)	Self-attention models relationships between image tokens across wider spatial ranges (Vaswani et al., 2017; Dosovitskiy et al., 2021)	Combines strong local feature extraction with wider contextual modelling.
Inductive bias / data efficiency	Strong locality bias; usually more data-efficient on moderate-sized image datasets (LeCun, Bengio and Hinton, 2015)	Weaker image-specific inductive bias (weaker built-in preference for image patterns such as local shapes and nearby pixels); often needs stronger regularisation or more data (Dosovitskiy et al., 2021)	CNN branch provides a stable feature foundation, while the transformer branch contributes broader context.
Compute scaling	Cost scales with image size, channels and network depth	Self-attention cost increases rapidly with token count	A hybrid can retain CNN efficiency while keeping transformer token usage controlled.
Handling orientation variation	Not explicitly rotation-aware by default	RViT-based design integrates information across rotated views (Krishnan et al., 2024)	The transformer branch is relevant where MRI orientation may vary across scans or preprocessing.
Interpretability hooks	Convolutional feature maps support class-localisation methods such as Grad-CAM and Grad-CAM++ (Selvaraju et al., 2017; Chattopadhyay et al., 2018)	Token-attention analysis such as rollout can provide transformer-side diagnostics, although attention alone is not a full explanation (Abnar and Zuidema, 2020; Jain and Wallace, 2019)	The hybrid allows branch-specific evidence checks rather than relying on a single explanation view.

2.7 Prior work motivating pathology-centred focus and token efficiency

Strong test accuracy alone does not show that a model will remain dependable in real clinical use. Models may later face scans from different hospitals, scanners, preprocessing routines, or patient groups, and performance can weaken even if benchmark results looked strong during development (Roberts et al., 2021; Varoquaux and Cheplygina, 2022).

One important reason is shortcut learning, where a model relies on spurious or unintended correlations instead of genuinely relevant features (Geirhos et al., 2020). In medical imaging, such cues can come from scanner or hospital differences, preprocessing artefacts, or background structure rather than the pathology itself. For instance, if tumour images often contain a repeated corner marker, the model may associate that cue with the tumour label instead of learning its real visual characteristics.

Lin et al. (2024) show that this problem appears in other medical-imaging tasks, including skin-lesion and fetal-ultrasound segmentation (automatic marking of the relevant region in the image).
Their examples show that models can rely on visible non-pathology markers such as calipers (measurement markers on the image), or on dataset-construction bias such as objects appearing near the image centre, confirming that shortcut learning is a wider medical-imaging reliability problem. Figure 2.5 illustrates these failure modes visually.

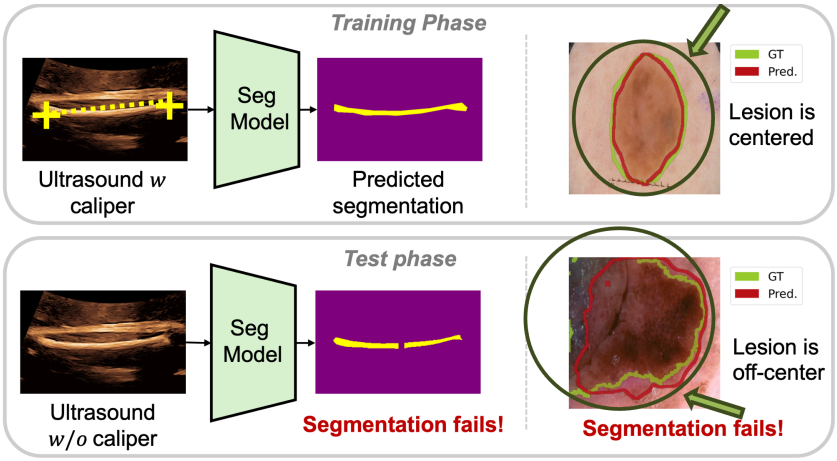


Figure 2.5: Conceptual illustration of two forms of shortcut learning in medical image analysis. On the **left**, a model may rely on visible non-pathology markers, such as calipers or annotations, and then fail when those markers are absent. On the **right**, the circled regions show both a lesion-centred case, where the arrow indicates successful segmentation, and an off-centre case, where the arrow indicates failed segmentation. Adapted from Lin et al. (2024, fig. 1, p. 2).

Although these studies do not propose a direct shortcut-learning solution for brain-tumour MRI classification, they offer ideas that can be adapted here. Xia, Chartsias and Tsaftaris (2020) show that pathology-related content can be separated from background content, which motivates learning a soft tumour-focused mask (a mask that gives regions different levels of importance rather than a simple on/off selection) from CNN features. Rao et al. (2021) show that not all transformer tokens are equally useful, and Bolya et al. (2023) show that similar tokens can be combined to reduce redundant processing.

In this project, these ideas are applied more simply through the proposed PFD-GSTE variants, which encourage greater focus on tumour evidence and less reliance on background cues, because brain MRI datasets can contain acquisition bias (differences caused by how the scan was taken) and position-related cues.

2.8 Explainability and uncertainty for human-centred evaluation

Healthcare AI evaluation should also show where the model focused and whether its confidence can be trusted. Although Gradient-weighted Class Activation Mapping (Grad-CAM) can in principle be applied to both branches in this project, it was not applied to both branches here. Instead, Grad-CAM++ was used for the CNN branch and Attention Rollout for the transformer branch.

In Figure 2.6(a), Grad-CAM++ is suited to CNNs because it uses gradients from convolutional feature maps to highlight image regions that support the predicted class. In Figure 2.6(b), Attention Rollout is better suited to transformers because it traces attention across layers to estimate which input tokens influenced the decision. Both produce heatmaps, but they represent different explanation types: Grad-CAM++ is feature-map based, while Attention Rollout is token-attention based. Using one method for the CNN branch and another for the transformer branch is therefore more appropriate than forcing a single explanation method on the whole hybrid. From a trust perspective, these views help show whether predictions appear tumour-centred or potentially superficial.

However, saliency maps can be unstable and attention weights are not explanations by themselves (Adebayo et al., 2018; Jain and Wallace, 2019). This project therefore also uses MC Dropout, where dropout remains active during inference and the same image is passed through the model multiple times to observe prediction stability or variation (Gal and Ghahramani, 2016). Since XAI shows where the model focuses but not how certain it is, combining both gives a more complete check of model behaviour.

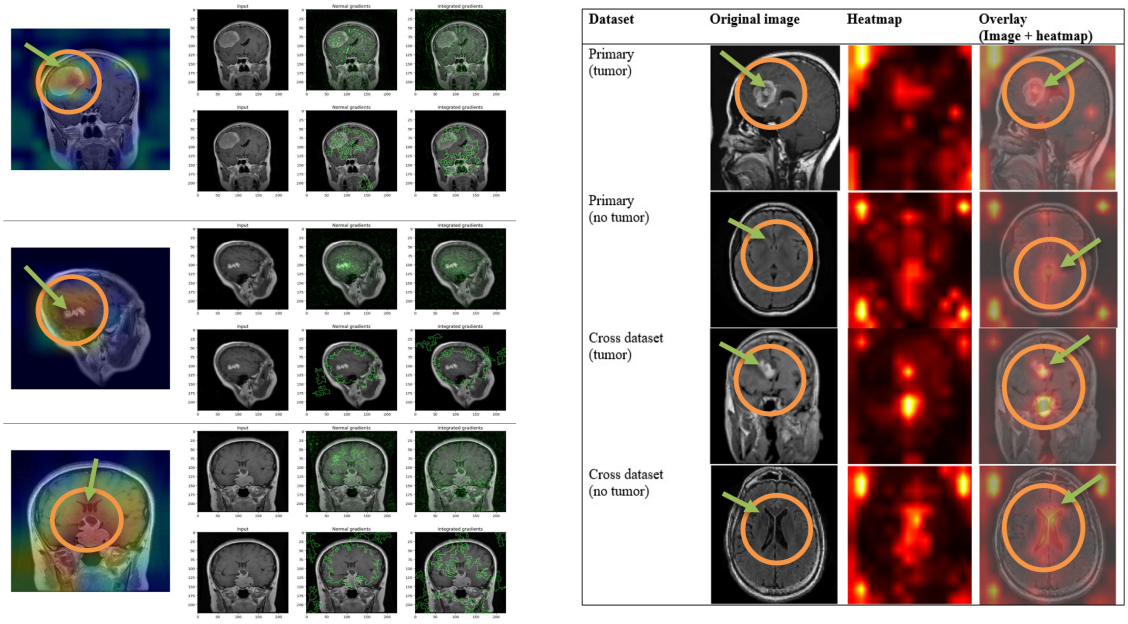


Figure 2.6: **Left (a)**: Example Grad-CAM++ heatmap over MRI showing class-relevant regions. Adapted from Sánchez-Moreno et al. (2025, fig. 6, p. 9). **Right (b)**: Example Attention-Rollout heatmaps for a ViT model showing the original image, heatmap, and overlay. Adapted from Tian (2025, fig. 6, p. 13). In both **(a)** and **(b)**, circled regions and arrows were added to point to the explainability regions highlighted by Grad-CAM++ and Attention Rollout.

2.9 Synthesis: gap and project justification

Taken together, the earlier sections show that prior studies support the separate value of the selected design components, but they do not directly test whether a guided hybrid model can reduce shortcut learning in a four-class brain tumour MRI setting. As summarised in Table 2.2, existing work mainly reports classification performance, with less attention to whether predictions rely on tumour-related evidence under strict development-test separation and combined explainability and uncertainty analysis.

This project addresses that gap by investigating a ResNet50V2-RViT hybrid with PFD-GSTE guidance. Reporting is guided by the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) to support transparent documentation of the dataset, preprocessing, splits, metrics, and reproducibility (Mongan, Moy and Kahn, 2020).

Table 2.2: Research gap synthesis matrix

Study	Main contribution	Limitation carried into this project
Sarada et al. (2025)	Modified ResNet50V2 CNN benchmark for brain-tumour MRI classification	CNN-only; shortcut-learning mitigation and branch-specific interpretability are not central
Karagoz et al. (2024)	Hybrid CNN-ViT design with strong classification motivation	Focus is mainly predictive performance, not shortcut-learning analysis under strict test-set separation with corresponding ablations
Krishnan et al. (2024)	Rotation-Invariant ViT (RViT) with interpretability support	Reported on binary classification, not the four-class subtype problem addressed here
This project	ResNet50V2-RViT hybrid with PFD-GSTE guidance, XAI, and uncertainty	Tests whether guidance improves tumour-focused evidence under strict test-set separation in a four-class evaluation

3. Project Method

This chapter describes the implementation procedure used in the project. It outlines the main stages of dataset preparation, split construction, hybrid model design, guidance-module integration, explainability, uncertainty estimation, and ablation. The overall workflow is summarised in Figure 3.1. The list of external software libraries, frameworks, and implementation tools used in the project is provided in Appendix A2.



Figure 3.1: Overall project workflow showing dataset checking, duplicate removal, train/validation/test split creation, fixed preprocessing, hybrid model design, training and evaluation of all model variants, and generation of explanation outputs.

3.1 Legal, ethical and professional considerations

This project is an experimental prototype, not a clinical decision tool; its outputs are used only for experimental evaluation, not for diagnosis or treatment. The main risks are misclassification and shortcut learning, where the model may rely on irrelevant image cues instead of the tumour region (Geirhos et al., 2020).

To reduce these risks, the project uses a public anonymised dataset with no names, identifiers, or patient metadata, applies duplicate handling to protect the held-out test set, and includes explainability outputs to check whether predictions appear tumour-focused. Only the minimum data needed are processed, no re-identification is attempted, access is limited to the project workspace, and the dataset is not redistributed beyond the original public source (University of Hertfordshire, 2021; Information Commissioner's Office, 2023; European Union, 2016).

3.1.1 Dataset source and provenance

The dataset used is the Brain *Tumor* MRI Dataset by Masoud Nickparvar (2021), hosted on Kaggle. It combines images from multiple sources, including figshare, SARTAJ, and Br35H, and provides four labelled classes: glioma, meningioma, pituitary, and no tumour (Nickparvar, 2021). Because the dataset is aggregated, audit checks, duplicate checking, and careful split control were needed before fair model comparison.
Nickparvar also notes that suspected glioma-class issues in the SARTAJ source were handled by using figshare images instead. This was treated as part of the dataset source record and as another reason to apply independent duplicate checking and careful split strategy before training.

3.2 Dataset description and label structure

The Kaggle dataset is organised into two top-level folders, Training and Testing, which are treated as an intended development/test separation rather than merged. The dataset contains 7023 images in total (Nickparvar, 2021). The class-to-integer mapping used consistently throughout preprocessing, training, evaluation, plotting, and explainability is shown in Table 3.1. Fixing this mapping across all scripts prevents label-order mismatches in metrics, confusion matrices, and visual outputs. The folder label *notumor* is referred to as no tumour in the report for readability.

Table 3.1: Class labels and integer mapping

Class name (folder label)	Meaning	Integer label
glioma	tumour type	0
meningioma	tumour type	1
pituitary	tumour type	2
notumor	no tumour present	3

3.2.1 Dataset curation and preprocessing pipeline

Dataset preparation is performed offline (performed before training) using a dedicated preprocessing script before any model training. The script converts the raw Kaggle folder structure into a processed tight-cropped dataset and split CSV files; training scripts later read only these prepared outputs, improving reproducibility and preventing run-to-run differences caused by accidental preprocessing changes. First, a raw manifest is created by scanning all folders and recording the file path, class label (derived from the folder name), and original Kaggle split (Training or Testing) for each image, creating an auditable (traceable) record of the dataset before filtering.

Next, exact duplicates are removed using SHA1 (Secure Hash Algorithm 1) hashes computed from image bytes (National Institute of Standards and Technology (NIST), 2015). If an image appears in both Kaggle Training and Kaggle Testing, the Testing copy is retained and the Training copy is removed so that duplicate test information cannot influence model development. Table 3.2 summarises the deduplication counts. Figure 3.2 supports this stage by showing the original Kaggle split distribution, the post-deduplication (after duplicate removal) class distribution, and an example preprocessing output.

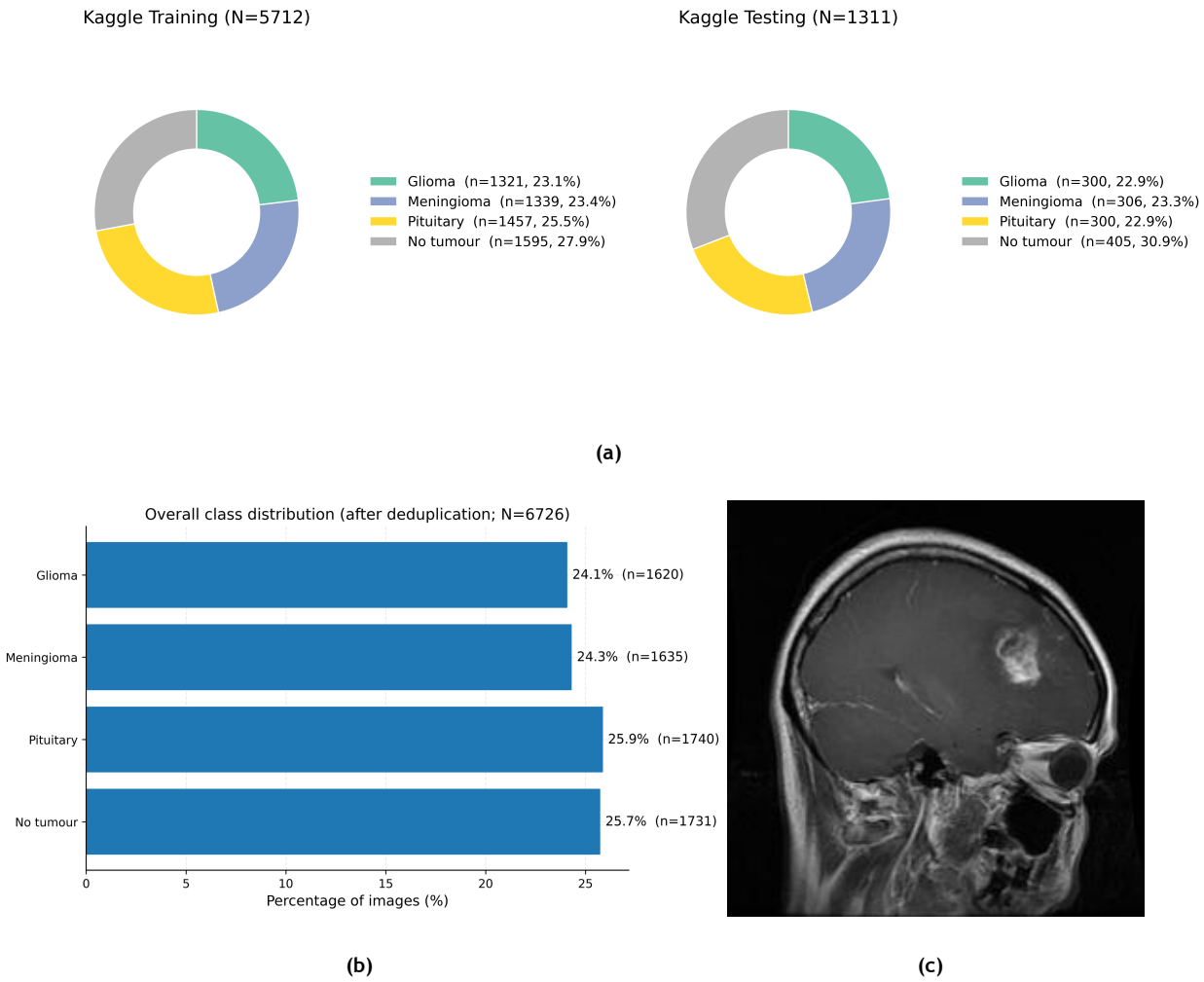


Figure 3.2: Dataset curation overview. **Upper-middle (a):** Original Kaggle Training/Testing split distribution (7023 images). **Lower-left (b):** Overall class distribution after SHA1 deduplication and duplicate handling with strict test-set protection. **Lower-right (c):** Example of preprocessing output after Exchangeable Image File Format (EXIF) correction, tight cropping, resizing to 224×224, and Red, Green, Blue (RGB) conversion.

Table 3.2: Deduplication results

Stage	Count
Raw images discovered	7023
Unique images after SHA1 deduplication	6726
Duplicate entries removed	297

After deduplication, each image is audited (checked carefully for quality and basic properties) for width, height, aspect ratio, and grayscale intensity statistics (mean, standard deviation, minimum, maximum). Conservative quality flags (too dark, too bright, low contrast, failed read) are computed to detect corrupted or abnormal inputs before training. Table 3.3 reports the audit summary, while Figure 3.3 provides representative examples inspected during this stage.

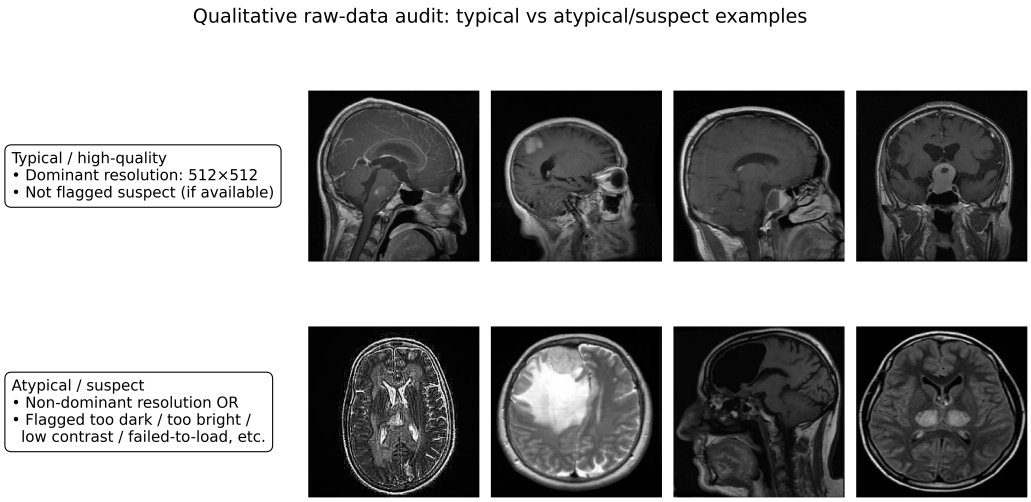


Figure 3.3: Representative examples reviewed during the raw-image audit, illustrating typical image quality and acquisition variability. No images were flagged as too dark, too bright, low contrast, or failed reads in the final audit summary (Table 3.3).

Table 3.3: Raw resolution and quality audit highlights

Audit output	Observation
Most common resolution	512 × 512 (4717 images)
Quality flags (too dark / too bright / low contrast / failed)	0 flagged in all classes
Total suspect images	0

Deterministic preprocessing (fixed steps) is then applied to all retained images: EXIF transpose (orientation safety), tight cropping to remove black background (threshold = 5, margin = 10), resizing to 224×224 using LANCZOS resampling, and conversion to RGB (Clark and contributors, 2024).

3.2.2 Data splitting and test-set protection

To preserve Kaggle's intended development/test separation while enabling validation-based model selection, Kaggle Training is divided into Train and Validation using a stratified 80/20 split (Pedregosa et al., 2011), while Kaggle Testing is preserved as the held-out Test set. This provides a validation set for early stopping and model selection without allowing test data to affect training decisions. The final per-class split counts are listed in Table 3.4, and Figure 3.4 shows the class distribution across the three splits. No augmentation is applied during this stage.

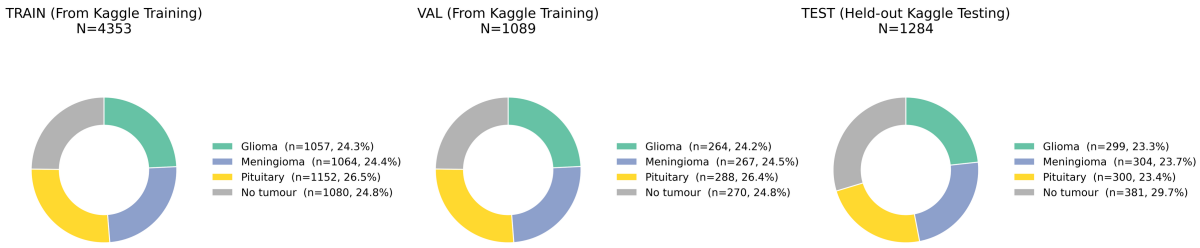


Figure 3.4: Final split distribution by class after stratified 80/20 Train/Validation division of Kaggle Training and preservation of Kaggle Testing as the held-out test set (Table 3.4).

Table 3.4: Final split counts per class

Split	glioma	meningioma	pituitary	notumor	Total
Train	1057	1064	1152	1080	4353
Val	264	267	288	270	1089
Test	299	304	300	381	1284

3.2.3 Dataset interface and online transformations

During model training, images are read only from the prepared split CSV files using the fixed class ordering defined in Table 3.1. The offline dataset remains deterministic; stochastic transformations are introduced only at training time to improve generalisation. In this project, light augmentation is applied to the training split only (small rotations, flips, translations, and optional low-magnitude noise), while validation and test inputs are not augmented. All splits are converted to tensors and normalised using the same channel statistics. The augmentation and loader settings used in all experiments are specified in Chapter 4.

3.3 Baselines, alternatives considered, and rationale

Sarada et al. (2025) describe a modified ResNet50V2 for brain-tumour MRI classification and report about 96.33% multi-class accuracy, but the paper does not provide enough layer-level detail for exact reproduction. This project therefore uses a reproducible pretrained PyTorch Image Models (timm) ResNet50V2 backbone (resnetv2_50x1_bitm) as the CNN feature extractor, returning the final 7×7 feature map for 224×224 input (Wightman, 2019). A small CNN head is then added within the hybrid design: Hybrid A uses global average pooling, linear projection, and rectified linear unit (ReLU), while

Hybrid B adds dropout and batch normalisation. This provides a stable CNN branch while allowing the effect of the guidance modules to be tested clearly. The transformer branch is motivated by Krishnan et al. (2024), whose RViT addresses feature variation under rotation. Hybrid A uses tokens from a guided CNN feature map, whereas Hybrid B follows the RViT idea more directly by rotating the raw image, extracting patch tokens, and integrating rotation-aware embeddings before transformer encoding.

3.3.1 Options considered but not addressed

Several extensions were considered but not included. Perceptual-hash near-duplicate detection was excluded because threshold selection and manual review would add uncertainty. Three-dimensional and multi-sequence MRI modelling were excluded because the chosen dataset consists of independent 2D images and does not provide reliable 3D continuity or aligned sequences. Segmentation-guided training was also excluded because tumour masks were not available and would have changed the project from a classification task into a joint segmentation-classification task.

3.4 System architecture: hybrid CNN-Transformer

The project implements two related hybrid architectures (Hybrid A and Hybrid B). In both variants, a CNN branch (ResNet50V2) extracts deep visual features, and a transformer branch (RViT-based) models wider spatial relationships using tokenised representations. The branch descriptors are combined by late fusion (concatenation followed by a small classifier) to produce four output logits.

3.4.1 Hybrid A

Hybrid A takes a preprocessed 224×224 RGB MRI image and passes it through ResNet50V2, producing a deep 7×7 feature map. From this map, the model first forms a CNN summary using global average pooling, a linear projection, and ReLU, giving a stable global representation of the image. PFD-A then learns a soft 7×7 mask from the same feature map. This mask gives more weight to regions that appear more tumour-related, and the masked feature map is passed to the transformer branch.

The masked feature map is rotated into four directions: 0°, 90°, 180°, and 270°. Each rotated 7×7 map is converted into 49 tokens, with one token per spatial cell. GSTE-A then reweights these tokens using the rotated mask values, so stronger regions receive more emphasis while the token count remains unchanged. Positional and rotation embeddings are then added. The token sets from the four rotations are averaged and passed through the RViT encoder. The encoder output is mean-pooled and projected into a transformer summary vector. Finally, the CNN summary and transformer summary are concatenated and passed through a fusion layer, dropout, and the final classifier to produce the four output logits (Figure 3.5).

3.4.2 Hybrid B

Hybrid B also begins with a preprocessed 224×224 RGB MRI image, which is passed through ResNet50V2 to produce a 7×7 feature map. PFD-B learns a soft mask from this map and applies it to the feature map. Unlike Hybrid A, this mask-guided feature map is used not only for transformer guidance but also for the CNN summary, so the CNN branch is pushed more directly towards likely tumour regions. The CNN summary is then formed using global average pooling, dropout, a linear projection, batch normalisation, and ReLU.
At the same time, the transformer branch operates on the raw image rather than CNN feature-map cells. The image is split into 16×16 patches, giving a 14×14 grid or 196 tokens. The learned mask is resized to this patch space and used as a guidance grid. GSTE-B reweights the tokens so that those from stronger regions receive more emphasis. When the mask is concentrated in a smaller area, the token grid can also be reduced by weighted pooling, allowing the model to focus more on the most relevant region instead of keeping all 196 tokens at full size.

As in Hybrid A, the image is processed in four rotations. For each rotation, patch tokens are created, reweighted by the rotated guidance grid, and given positional and rotation embeddings. The token sets are averaged across rotations and passed through the RViT encoder. The encoder output is mean-pooled and projected into a transformer summary vector. Finally, the CNN summary and transformer summary are fused and passed to the classifier to produce the four output logits (Figure 3.6).

Most labels in Figures 3.5 and 3.6 match the implementation, although some are shortened for clarity. GAP, FC and logits denote global average pooling, fully connected layer, and raw class scores before softmax. In Hybrid B, dyn_side is the dynamically selected token-grid side length after mask-guided shrinking, and α is the mask-derived token weight. The variables feat, mask_feat, feat_path, z_cnn, z_vit, and Tavg refer to the CNN feature map, learned mask, mask-guided feature map, CNN summary, transformer summary, and the token set averaged across the four rotations.
Cubes with grids represent spatial feature maps, flat slabs represent transformer tokens, and repeated slabs indicate rotated versions before combination.

3.4.3 Architectural comparison

The key difference is that Hybrid A uses 49 CNN-derived tokens and applies guidance mainly to the transformer pathway, whereas Hybrid B uses up to 196 image-patch tokens and applies guidance to both the CNN and transformer pathways. Hybrid A therefore represents the weaker guidance design, while Hybrid B represents the stronger one.

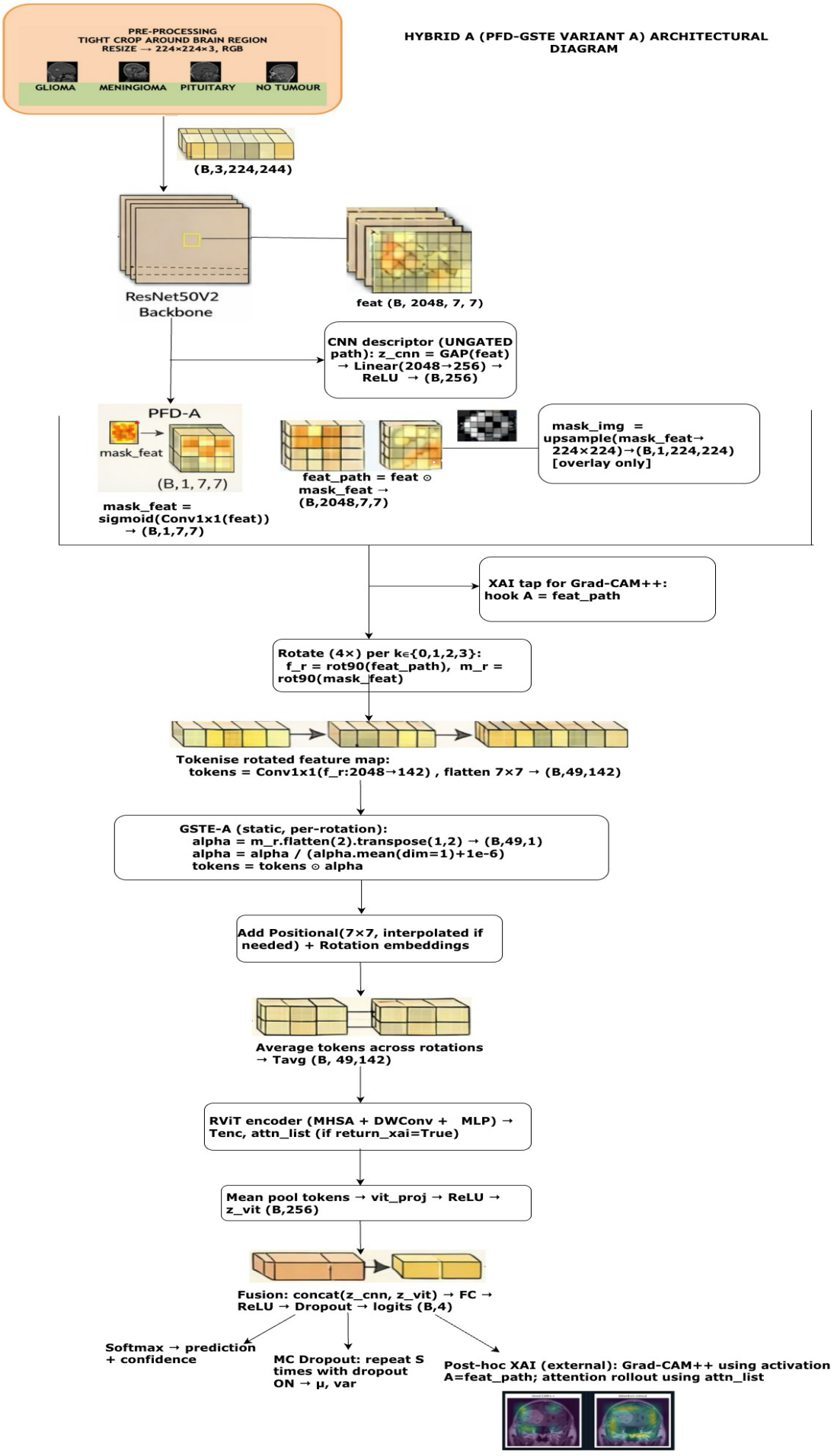


Figure 3.5: Hybrid A architecture (CNN-feature-token transformer) with PFD-A/GSTE-A guidance.

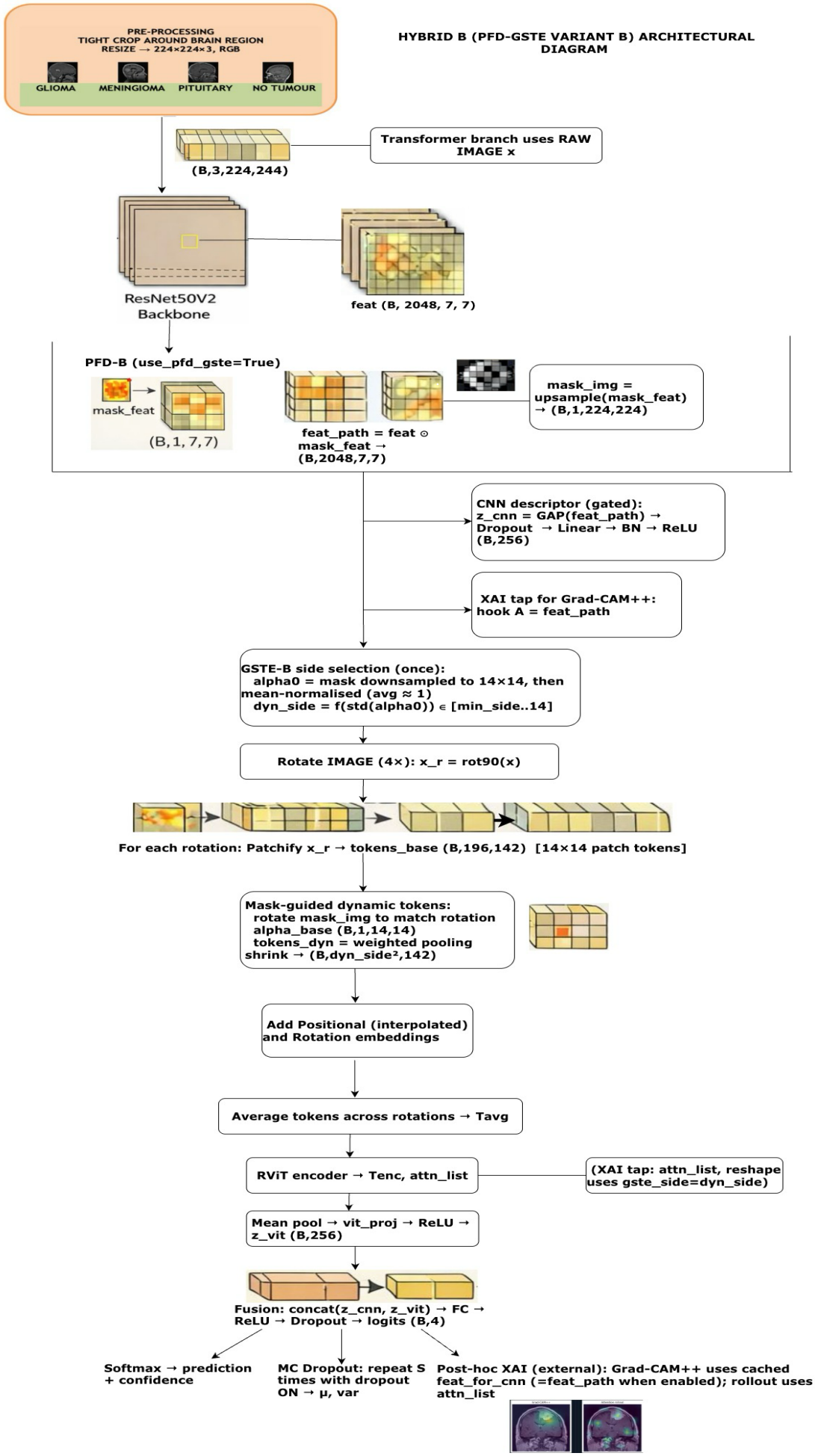


Figure 3.6: Hybrid B architecture (patch-token transformer) with PFD-B/GSTE-B guidance and optional mask-guided token-grid shrinking.

3.4.4 Proposed contributions: PFD and GSTE

Building on Section 2.7, this project introduces two experimental guidance modules into the ResNet50V2-RViT hybrid: *Pathology-Focused Disentanglement* (PFD) and *Guided Semantic Token Evolution* (GSTE). These project-specific modules were inspired by prior work on pathology emphasis and token importance, rather than reproduced directly from any single method (Xia, Chartsias and Tsaftaris, 2020; Rao et al., 2021; Bolya et al., 2023). In this project, PFD provides a learned tumour-focused spatial mask and GSTE uses that guidance to steer transformer tokens.

Variant A applies weaker guidance and Variant B applies stronger guidance as deliberate experimental design choices to test whether weaker or stronger guidance is more effective for mitigating shortcut learning. Together, these variants allow the study to examine whether increasing guidance strength improves tumour-centred evidence, classification behaviour, both, or neither.

Table 3.5: PFD variants

Variant	Definition	Where applied in this project	Practical effect
PFD-A	Learns a soft mask from the CNN feature map and gates the feature map with it	Gates the feature map used to create transformer tokens; CNN pooled descriptor remains ungated (left unchanged by the mask)	Nudges transformer reasoning toward tumour while keeping the CNN summary global and stable
PFD-B	Same learned mask and gating operation, used more strongly	Gates the CNN feature map for the CNN descriptor (the pooled summary vector from the CNN branch) and provides the mask used for transformer guidance	Forces the CNN summary to rely on pathology-focused activations, and encourages the transformer toward tumour-focused regions via mask-guided patch tokens (GSTE)

Table 3.6: GSTE variants

Variant	Definition	Token behaviour in this project	Purpose
GSTE-A	Mask-guided token weighting	Fixed token count (49 feature-map tokens); tokens are reweighted using mean-normalised mask weights	Simple, stable encouragement toward tumour evidence
GSTE-B	Mask-guided token weighting and optional token-grid shrinking	Starts from 196 patch tokens (14×14); tokens are weighted by the pooled mask; when the mask is concentrated, tokens can be compressed to a smaller grid via weighted pooling	Concentrates modelling capacity on tumour regions and reduces background tokens

3.5 RViT branch implementation and relation to Krishnan et al. (2024)

Both Hybrid A and Hybrid B use the main RViT idea of processing each MRI at four orientations (0°, 90°, 180°, and 270°) and adding rotation-specific embeddings before transformer encoding. In both variants, the transformer output is obtained by global token pooling rather than by requiring a classification token (CLS token).

Hybrid A applies this idea in a more adapted form, because its transformer branch operates on 49 guided CNN-derived feature tokens rather than raw-image patches. Hybrid B is closer to Krishnan et al. (2024), because it rotates the raw image, forms patch tokens, adds rotation information, and then passes them through the RViT-based encoder. These design choices were kept to adapt the RViT idea to the project's hybrid setting. They allowed the transformer branch to remain rotation-aware while fitting the ResNet50V2 fusion design.

3.5.1 Fusion of ResNet50V2 and RViT branches

As introduced in Section 3.4, both hybrids use late fusion, meaning the CNN and transformer first produce separate summaries and are merged only near the classifier. This was chosen to keep local CNN evidence and wider transformer context separate until the end, which also makes branch-specific explanation easier. The fusion is defined as:

$$z = [z_{cnn}; z_{vit}]$$
$$h = \mathrm{ReLU}(W_f z + b_f)$$
$$\hat{y} = W_o(\mathrm{Dropout}(h)) + b_o$$

(Equation 1)

Here, z_{cnn} is the pooled CNN summary vector, z_{vit} is the pooled transformer summary vector, and $[z_{cnn}; z_{vit}]$ means these two vectors are joined into one longer vector. W_f and b_f are the weights and bias of the fusion layer, ReLU is the activation function, Dropout reduces overfitting, and W_o and b_o map the fused representation to the four output logits. $\hat{y}$ is the final output before softmax.
Earlier fusion was not used because this study required clear and consistent branch separation for fair comparison of the PFD-GSTE variants (Yadav, Kumar and Nandi, 2025; Garcia-Atutxa et al., 2025).

3.6 Explainability integration

Explainability is applied after prediction and is not part of training. During early development, a single Grad-CAM approach was tried across the full hybrid, but it did not work reliably for these hybrid models because the heatmaps were noisy and the method did not match the model structure well. The final design therefore uses Grad-CAM++ for the CNN branch and Attention Rollout for the transformer branch (Selvaraju et al., 2017; Chattopadhyay et al., 2018; Abnar and Zuidema, 2020).

In the guidance-enabled models, Grad-CAM++ is computed from the PFD-modified feature map captured by a forward hook in the XAI script, producing a class-specific heatmap over the CNN evidence pathway. Attention rollout is computed from the transformer attention matrices returned during the forward pass. As described in Section 3.4, Hybrid A uses a coarse 7×7 token grid, so its rollout appears spatially coarser and less detailed, whereas Hybrid B begins from image patches and can retain a more detailed spatial pattern. Together, these two views support qualitative inspection of whether predictions appear tumour-centred or potentially driven by irrelevant cues.

3.6.1 Uncertainty estimation

To complement explainability, test-time uncertainty was estimated using MC Dropout (Gal and Ghahramani, 2016). This was done by passing the same MRI through the trained model multiple times while keeping dropout active during inference. The model was first set to evaluation mode so that Batch Normalisation (BatchNorm) used its stored statistics, and only the dropout layers were then turned back on. By default, 20 stochastic forward passes were used. Because dropout randomly removes part of the network activity on each pass, the predicted class probabilities vary slightly across runs.

These outputs were collected at the final four-class classifier and summarised. The mean probability was used as the final confidence estimate, while the variance was used as the uncertainty signal. In Hybrid A, most variation comes from dropout in the transformer and final fusion layers; in Hybrid B, it also comes from dropout in the CNN descriptor pathway before fusion. This is useful in the project because the heatmaps show where the model appears to focus, while MC Dropout shows how stable that prediction remains across repeated stochastic passes.

3.7 Ablation study and supporting web interface

The ablation study was used to test whether PFD and GSTE actually changed model behaviour. In this project, Ablation A is Hybrid A without PFD-A and GSTE-A, and Ablation B is Hybrid B without PFD-B and GSTE-B. Apart from removing these guidance modules, all other procedures remained unchanged.

A supporting prototype web interface was also developed for qualitative inspection of the trained models. It was not part of the main benchmark evaluation; further details are provided in Appendices A3 and A4.

4. Training

This chapter describes the shared training and model-selection procedure used for all four variants. Only the fixed dataset splits defined in Chapter 3 were used, so later performance differences can be interpreted primarily in relation to model design rather than changes in data partitioning.

4.1 Training objective and comparisons

The training objective was to learn a robust four-class MRI classifier (glioma, meningioma, pituitary, and no tumour) and to test whether PFD-GSTE guidance changed performance or evidence quality. Four configurations were trained: Hybrid A, Hybrid B, Ablation A, and Ablation B.

Training used single-device selection with Compute Unified Device Architecture (CUDA) if available, otherwise central processing unit (CPU), and was conducted in the Kaggle environment. Reproducibility was supported by fixing the random seed (seed = 42) for NumPy and PyTorch, while CUDA Deep Neural Network library (cuDNN) (the GPU library used to speed up deep-learning operations) benchmarking remained enabled for faster kernel selection (automatic choice of the fastest way to run an operation). This improved consistency across runs but did not enforce strict deterministic execution. Hyperparameter settings for all four variants are listed in Appendix A
(Tables A.1.1-A.1.2). Data loading was consistent across runs: training data were shuffled, validation and test were not, and the DataLoader used num_workers=2 and pin_memory=True.

4.1.1 Training-time input processing: augmentation and normalisation

As outlined in Section 3.2.3, the offline dataset remains deterministic, and stochastic variation is introduced only during training. The training split used light augmentation: random rotation ($\pm$15°), random horizontal flip (probability = 0.5), small image shifts (up to 5% of image width/height), and optional low-magnitude Gaussian noise (std = 0.02, p = 0.5).
After augmentation, images were converted to tensors and normalised using per-channel mean and standard deviation of (0.5, 0.5, 0.5). Validation and test images were not augmented; they were only tensor-converted and normalised with the same channel statistics.

This defined the common input distribution for all four variants. Their effect on optimisation stability is examined in Section 4.2.

4.1.2 Optimisation strategy

All four models used the same optimisation procedure. Training ran for up to 100 epochs with a batch size of 32. The optimiser was Adaptive Moment Estimation with decoupled weight decay (AdamW) with differential learning rates: 1×10^{-4} for the pretrained CNN backbone and 5×10^{-4} for the transformer and fusion components. Weight decay was 0.01, with no decay applied to biases, normalisation parameters, or other one-dimensional parameters. Training used cross-entropy loss with label smoothing of 0.05, while validation and test reporting used standard cross-entropy. Because these losses differ, their values are not directly comparable, so checkpoint selection relied on validation macro-F1 rather than loss alone.

A cosine annealing learning-rate schedule was used with $\eta_{\min} = 1 \times 10^{-6}$. The CNN backbone was frozen for the first five epochs and then unfrozen for fine-tuning. At that point, the optimiser and scheduler were rebuilt so the schedule covered the remaining epochs correctly. This helped keep outcome differences mainly attributable to model design rather than extra tuning.

4.1.3 Stability controls and implementation details

To improve optimisation stability, gradients were clipped to a maximum norm of 1.0. The implementation also supports optional Automatic Mixed Precision (AMP) (a method that uses lower-precision calculations to reduce GPU memory use and speed up training) and optional freezing of CNN BatchNorm statistics (keeping the stored normalisation values fixed instead of updating them during training); neither was used in the reported experiments (Micikevicius et al., 2018).

One small implementation difference remained to preserve each family's implemented training setup: Hybrid B and Ablation B used drop_last=True, so the final smaller training batch was dropped, whereas Hybrid A and Ablation A used the default drop_last=False, so the final smaller batch was kept.

4.1.4 Model selection and checkpoint saving

Model selection was performed only on the validation split using early stopping on validation macro-F1 with a patience of 10 epochs. Macro-F1 was chosen because it weights all classes equally and reduces the chance of favouring a model mainly because one class is easier or more frequent. The selected checkpoint was the epoch with the best validation macro-F1.

The best checkpoint was saved with the model state, class ordering, normalisation values, training arguments, and model configuration needed for later reconstruction at inference time.

Table 4.1: Model selection outcomes (validation-based early stopping)

Model variant	Best epoch	Best val macro-F1
Hybrid A	43	0.9927
Hybrid B	14	0.9880
Ablation A	30	0.9936
Ablation B	42	0.9945

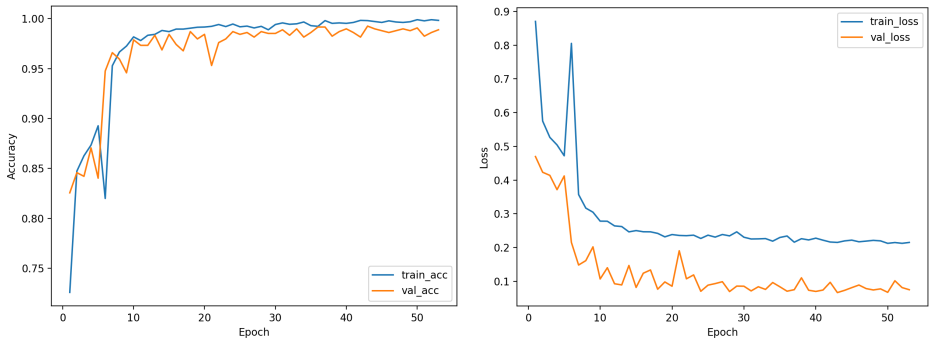
4.1.5 Evaluation metrics and reporting

The reported metrics are standard classification measures, selected to summarise both overall performance and the pattern of classification errors (Sarada et al., 2025; Krishnan et al., 2024).

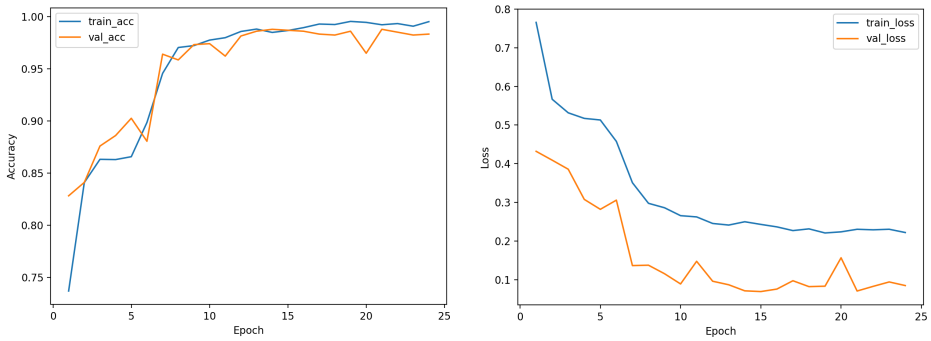
- Accuracy (Acc): overall proportion of test images classified correctly.
- Cross-entropy loss (Loss): how wrong the predicted probabilities are; lower is better.
- Precision: of the images predicted as a class, how many were correct.
- Recall / Sensitivity: of the images truly in a class, how many were found correctly.
- F1-score: a balance between precision and recall.
- Macro-F1 (MF1): average F1 across all classes, giving each class equal importance.
- Weighted-F1 (WF1): average F1 across all classes, weighted by class size.
- Specificity: how well the model correctly rejects other classes.
- Cohen's kappa (κ) and MCC: summary measures of overall classification agreement.

4.2 Training dynamics and convergence

The learning curves show that all four variants converged successfully under the shared optimisation procedure described in Sections 4.1.2-4.1.3. Figure 4.1 shows the accuracy and loss curves for the full hybrid variants, and Figure 4.2 shows the corresponding curves for the ablation variants. Overall, all runs improved quickly in the early epochs and then reached high validation performance. Severe persistent overfitting is not evident, although some later validation fluctuations remain after convergence.

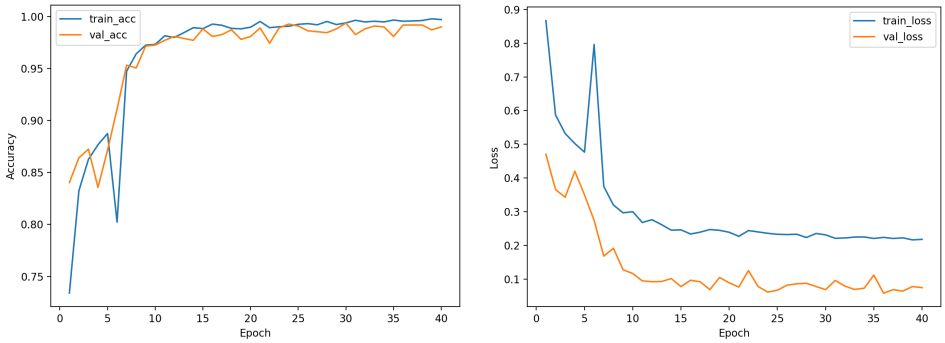


(a) (b)

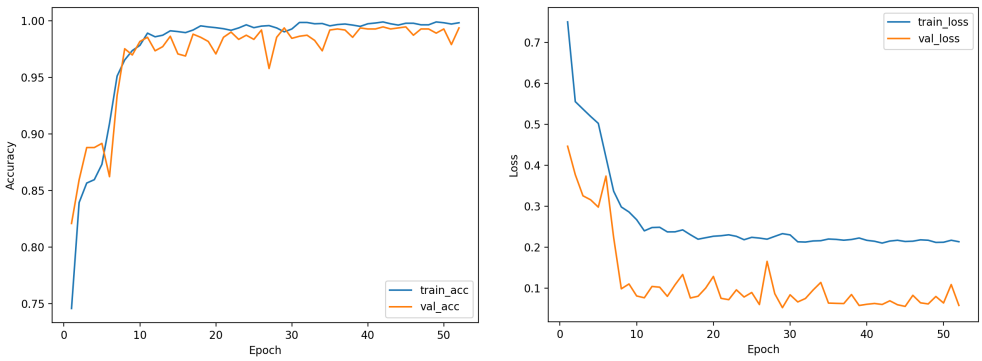


(c) (d)

Figure 4.1: Training dynamics for the full hybrid variants. **Upper Left (a):** Hybrid A accuracy, **Upper Right (b):** Hybrid A loss, **Lower Left (c):** Hybrid B accuracy, and **Lower Right (d):** Hybrid B loss, each showing training and validation curves under the shared optimisation protocol (AdamW with differential learning rates, cosine annealing, and 5-epoch freeze-unfreeze warmup).



(a) (b)



(c) (d)

Figure 4.2: Training dynamics for the ablation variants. **Upper Left (a):** Ablation A accuracy, **Upper Right (b):** Ablation A loss, **Lower Left (c):** Ablation B accuracy, and **Lower Right (d):** Ablation B loss, each showing training and validation curves under the same shared optimisation protocol.

The fluctuation points below were identified from the saved training history files and match the trends shown in Figures 4.1 and 4.2.

- Hybrid A (Figure 4.1a-b): Training and validation performance remained strong overall. Small later fluctuations appeared around epochs 21, 34, 42, and 52, showing mild late-stage instability rather than a major training problem.

- Hybrid B (Figure 4.1c-d): Best validation performance was reached around epochs 14-15. After this, validation performance dropped slightly at some points, especially around epochs 20, 23, and 24. This was the clearest case of mild late-stage instability and supports the use of early stopping.

- Ablation A (Figure 4.2a-b): After a short early fluctuation near the warmup transition, the model converged well. Later changes around epochs 22, 31, and 35 suggest mild temporary instability rather than clear overfitting.

- Ablation B (Figure 4.2c-d): Validation performance remained strong overall, with short fluctuations around epochs 15-16, 20, 27, 34, 46, and 52. These again suggest mild temporary instability rather than serious overfitting.

Overall, the curves suggest that optimisation was stable and that validation macro-F1 was a more suitable checkpoint-selection measure than loss alone.

5. Results

This chapter reports held-out test results for the four trained variants, including classification metrics, error patterns, and branch-specific explainability and uncertainty outputs.

Kaggle Testing was kept fully separate from model development and used only for final held-out evaluation. Each variant was evaluated once on this test set using its single best checkpoint selected by validation macro-F1. The test set contains 1284 images: glioma (299), meningioma (304), pituitary (300), and no tumour (381) (Table 3.4).

5.1 Overall held-out test performance

Table 5.1 summarises held-out test performance for all variants. All models achieved high held-out accuracy (0.9852-0.9922). Numerically, Ablation B achieved the highest accuracy and macro-F1 and the lowest loss, while Hybrid A and Ablation A were near-identical overall. Here, Params refers to the learnable weights and biases updated during training. All metrics were computed on the held-out test set from the exported evaluation summary of the selected checkpoint and are rounded to 4 decimal places.

Table 5.1: Held-out test performance across model variants.

Model variant	Test Acc	Test Loss	Test MF1	Test WF1	Spec (macro)	κ	MCC	Params
Hybrid A	0.9875	0.0803	0.9875	0.9875	0.9959	0.9833	0.9833	26.68M
Hybrid B	0.9852	0.0753	0.9849	0.9852	0.9952	0.9802	0.9803	26.58M
Ablation A	0.9875	0.0847	0.9873	0.9875	0.9959	0.9833	0.9834	26.68M
Ablation B	0.9922	0.0668	0.9920	0.9922	0.9975	0.9896	0.9896	26.58M

5.1.1 Per-class performance

Overall accuracy can hide class-specific weaknesses. Tables 5.2a-5.2d report per-class precision, recall (sensitivity), F1, and specificity. Across variants, the no tumour class remained consistently strong, while most remaining errors occurred at tumour subtype boundaries, as also seen in Figure 5.1 and Section 5.1.3.

Table 5.2a: Per-class metrics for Hybrid A

Class	Precision	Recall (Sensitivity)	F1	Specificity
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

Table 5.2b: Per-class metrics for Hybrid B

Class	Precision	Recall (Sensitivity)	F1	Specificity
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

Table 5.2c: Per-class metrics for Ablation A

Class	Precision	Recall (Sensitivity)	F1	Specificity
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

Table 5.2d: Per-class metrics for Ablation B

Class	Precision	Recall (Sensitivity)	F1	Specificity
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

Overall, Ablation B recorded the highest per-class values across most metrics, while Hybrid B showed lower glioma and meningioma precision.

5.1.2 Confusion matrices

Confusion matrices show the direction of remaining errors rather than only how many errors occur. Figure 5.1 presents held-out test confusion matrices for (a) Hybrid A, (b) Hybrid B, (c) Ablation A, and (d) Ablation B.

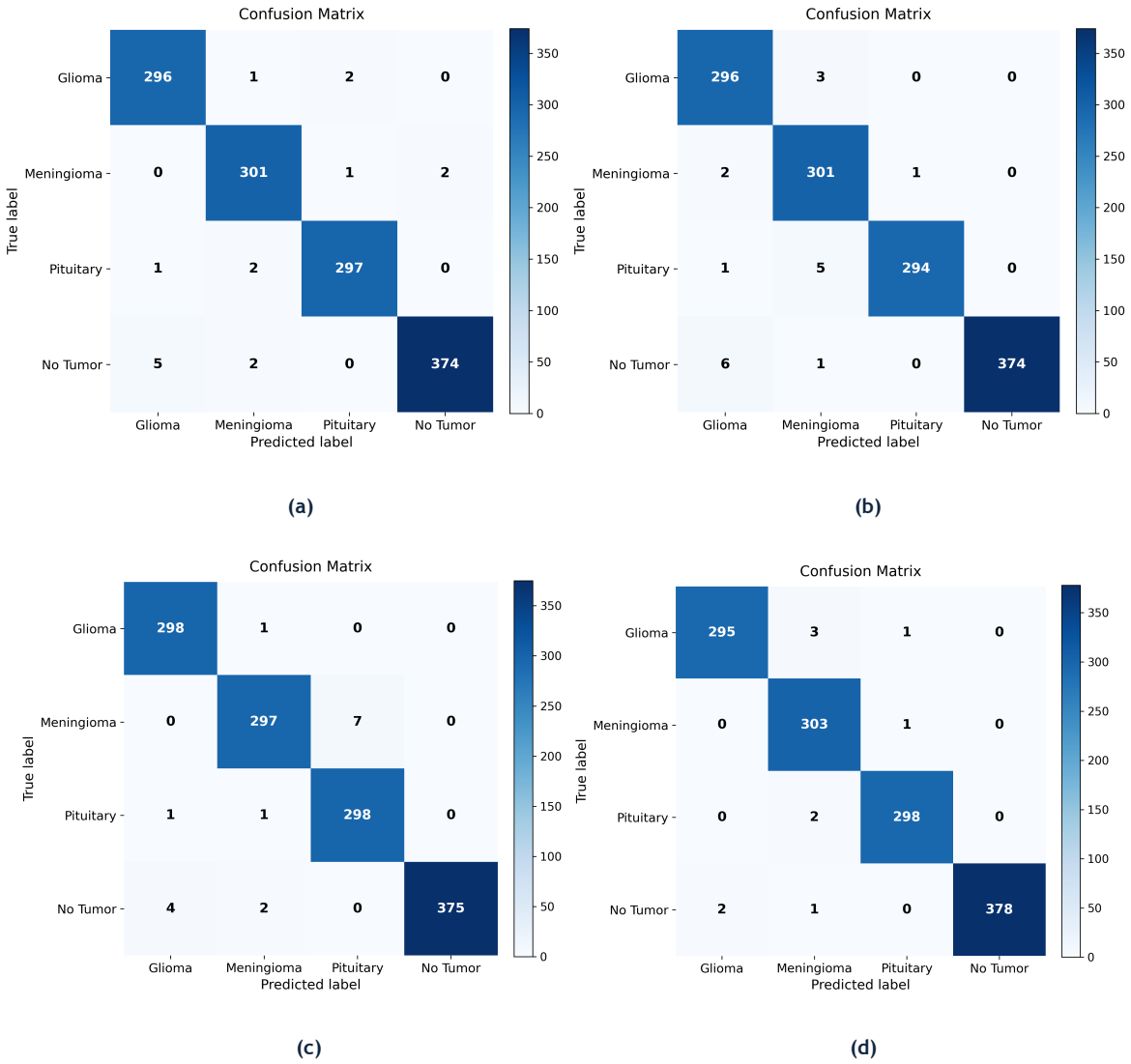


Figure 5.1: Held-out test confusion matrices for all model variants.
Upper Left (a): Hybrid A, **Upper Right (b):** Hybrid B, **Lower Left (c):** Ablation A, and **Lower Right (d):** Ablation B. Counts are shown for true classes (rows) against predicted classes (columns).

5.1.3 Error analysis with misclassification counts

The confusion matrices show that the remaining errors were concentrated mainly at tumour subtype boundaries rather than at the broad tumour-against-no tumour decision. Hybrid A and Ablation A each made 16 errors, but their patterns differed: Hybrid A produced more no tumour false positives, whereas Ablation A showed a stronger meningioma-to-pituitary confusion. Hybrid B produced the highest error count (19), with notable pituitary-to-meningioma and no tumour-to-tumour errors. Ablation B produced the fewest errors (10) and the most balanced error pattern overall.

Across all four variants, the dominant remaining problem was subtype confusion, with a smaller number of no tumour false positives.

5.1.4 Compact ablation difference summary

Table 5.3 summarises family-level deltas between each full hybrid and its ablation. Positive values favour the full hybrid; negative values favour the ablation. This table provides a bridge between Table 5.1 and the error counts above.

Table 5.3: Compact ablation difference summary (family-wise)
(Positive Δ Acc / Δ MF1 favours the full hybrid; negative values favour the ablation.)

Family	Full hybrid against ablation	ΔTest Acc (full − ablation)	ΔTest MF1 (full − ablation)	Error count (full against ablation)	Immediate reading
A-family	Hybrid A against Ablation A	0.0000	+0.0002	16 against 16	Numerically almost unchanged
B-family	Hybrid B against Ablation B	-0.0070	-0.0071	19 against 10	Ablation B performs better numerically

5.2 Explainability (XAI) results: Grad-CAM++ and Attention Rollout

This section reports the explainability outputs generated after training from the selected checkpoint of each model. Two views are used: Grad-CAM++ for the CNN branch and Attention Rollout for the transformer branch. The aim is to compare whether the highlighted region overlaps the tumour or shifts elsewhere in the image. The localisation labels were assigned by consistently comparing each heatmap with the visible tumour region marked in the figure, using the same judgement rule across all four variants.

5.2.1 Comparison setup

The same held-out test examples are compared across all four variants for consistency. The reported set includes one correct meningioma case, one challenging glioma case, and one pituitary-to-meningioma error case. This allows comparison between class prediction and visual focus. A correct class prediction does not always mean that the highlighted region is also correct.

The localisation labels were defined as Correct, Partially correct, Partially wrong, and Wrong. Partially correct was used when the highlighted region was mainly on the tumour with some minor non-tumour activation, whereas Partially wrong was used when only a small part of the tumour was highlighted and the main activation fell on non-tumour background.

5.2.2 Correct-case comparison

Figure 5.2 shows a held-out meningioma image classified correctly by all four models. It is used to compare whether correct classification is also supported by tumour-centred visual evidence. Although the class output is correct in every case, the highlighted regions differ in accuracy across Grad-CAM++ and Attention Rollout.

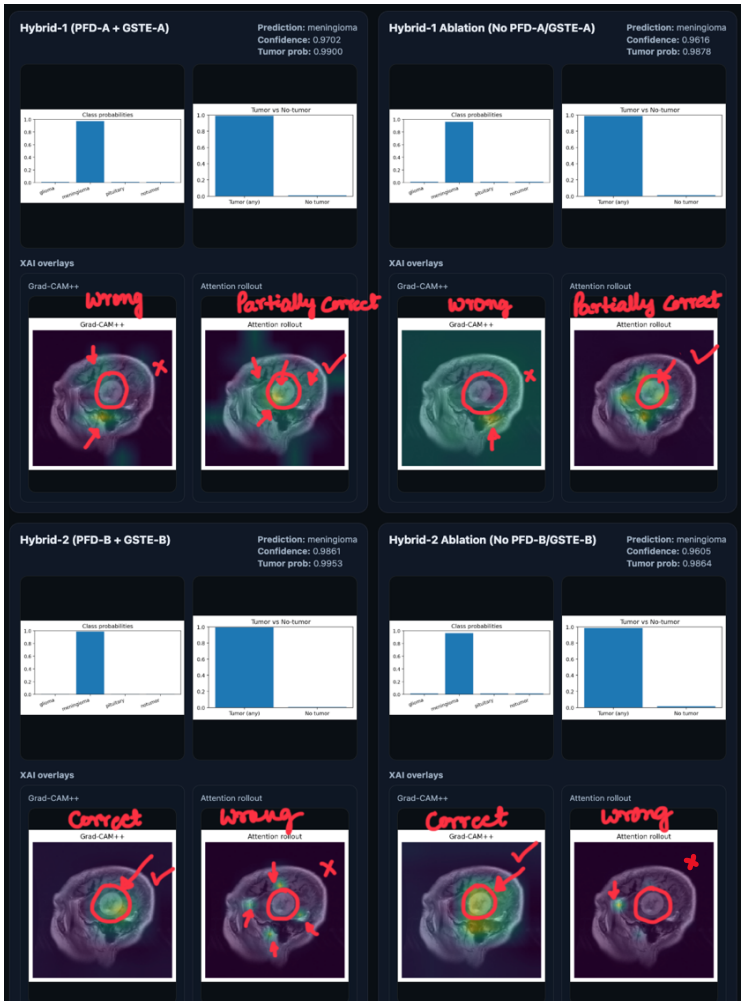
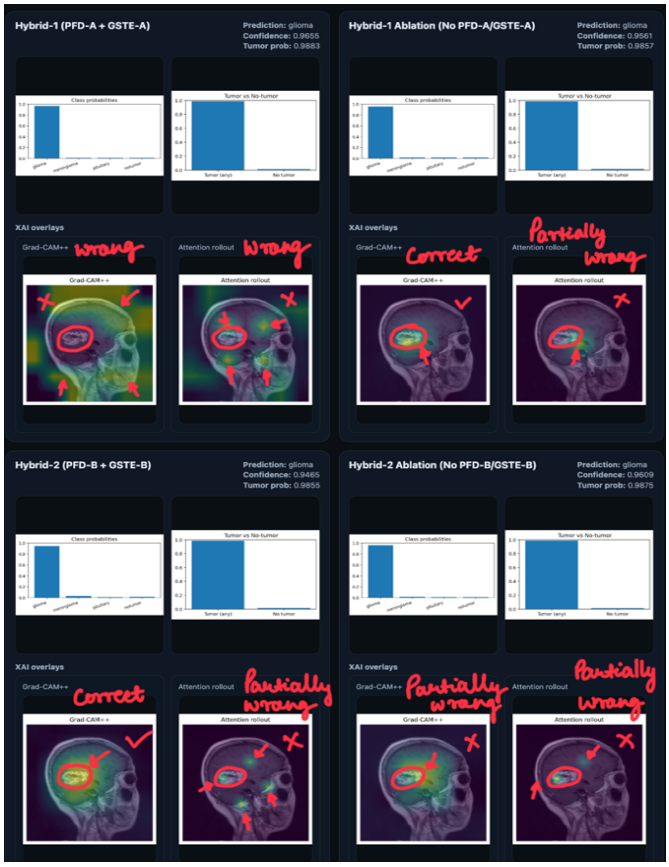


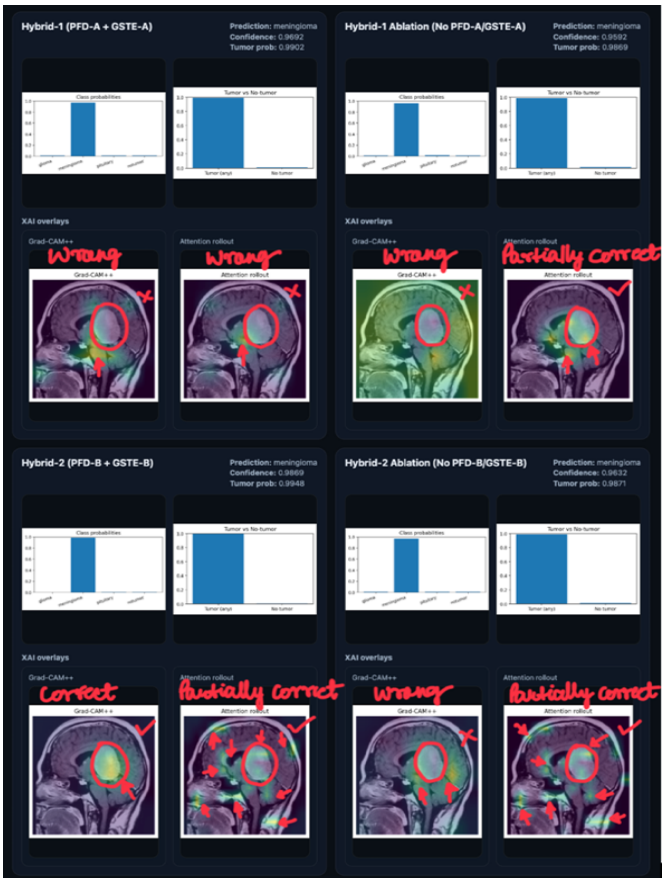
Figure 5.2: Correct-case explainability comparison for a held-out meningioma image. Red circles mark the visible tumour region and arrows indicate the main highlighted area. All four models predict the correct class, but localisation quality differs across Grad-CAM++ and Attention Rollout.

5.2.3 Challenging-case comparison

In this section, two more challenging held-out test cases are presented. Figure 5.3(a) uses a glioma example to compare localisation quality when the visual explanation is weak, shifted, or only partly aligned with the visible tumour region. Figure 5.3(b) presents a pituitary example misclassified as meningioma by all four models.



(a)



(b)

Figure 5.3: Challenging-case explainability comparison on held-out test images. Red circles mark the visible tumour region and arrows indicate the main highlighted area. **Upper (a):** Glioma example used to compare localisation quality when the prediction alone does not fully show where the model focused. **Lower (b):** Pituitary example misclassified as meningioma by all four models, showing how visual focus changes when the class prediction is wrong.

5.2.4 Visual and confidence summary

Table 5.4: Summary of explainability and confidence across the reported cases.
GC++ = Grad-CAM++; Roll = Attention Rollout; MC conf = Monte Carlo confidence; Tumour prob = summed tumour-class probability against no tumour.

Case	Main observation	Hybrid A	Ablation A	Hybrid B	Ablation B
Figure 5.2 Correct meningioma	All models predict the class correctly, but visual focus differs	GC++: Wrong; Roll: Partially correct; MC conf: 0.9702; Tumour prob: 0.9900	GC++: Wrong; Roll: Partially correct; MC conf: 0.9616; Tumour prob: 0.9878	GC++: Correct; Roll: Wrong; MC conf: 0.9861; Tumour prob: 0.9953	GC++: Correct; Roll: Wrong; MC conf: 0.9605; Tumour prob: 0.9864
Figure 5.3(a) Challenging glioma	Localisation quality differs across models	GC++: Wrong; Roll: Wrong; MC conf: 0.9655; Tumour prob: 0.9883	GC++: Correct; Roll: Partially wrong; MC conf: 0.9561; Tumour prob: 0.9857	GC++: Correct; Roll: Partially wrong; MC conf: 0.9465; Tumour prob: 0.9855	GC++: Partially wrong; Roll: Partially wrong; MC conf: 0.9609; Tumour prob: 0.9875
Figure 5.3(b) Pituitary misclassified as meningioma	All models are wrong, but confidence remains high	GC++: Wrong; Roll: Wrong; MC conf: 0.9692; Tumour prob: 0.9902	GC++: Wrong; Roll: Partially correct; MC conf: 0.9592; Tumour prob: 0.9869	GC++: Correct; Roll: Partially correct; MC conf: 0.9869; Tumour prob: 0.9948	GC++: Wrong; Roll: Partially correct; MC conf: 0.9632; Tumour prob: 0.9871

5.3 Uncertainty and confidence behaviour

To complement the visual explanations, predictive uncertainty was estimated using MC Dropout, as described in Section 3.6.1. In Table 5.4, MC confidence denotes the mean predicted probability for the selected class across repeated stochastic forward passes, while tumour probability denotes the summed probability assigned to the tumour classes relative to no tumour.

Across the reported cases, confidence often remained high even when localisation quality weakened and, in some cases, when the class prediction was incorrect. The pituitary-to-meningioma example was the clearest case, as all four models gave a wrong prediction while still producing high confidence.

6. Discussion and Evaluation

This chapter interprets the results against the project aim, focusing on whether the guidance modules improved evidence quality without harming overall classification performance.

6.1 Main outcome

The results indicate partial success. The clearest comparisons are within families rather than across families, because Hybrid A and Hybrid B differ not only in guidance strength but also in token design and one training-loader detail (drop_last). Within the A-family, Hybrid A and Ablation A were almost identical numerically, which suggests that the weaker guidance design had little effect on final classification. This is consistent with the architecture, because PFD-A and GSTE-A mainly influence the transformer-facing pathway, while the ungated CNN descriptor can still dominate the late-fusion decision.

Within the B-family, Ablation B achieved the best classification results, while Hybrid B showed the most tumour-centred Grad-CAM++ results in the reported cases. This suggests a trade-off:
the stronger guidance improved localisation more clearly than subtype classification. A likely reason is that PFD-B and GSTE-B concentrated attention more strongly on suspicious regions, which helped visual focus but may also have reduced some surrounding context needed for subtype separation. The stronger numerical result of Ablation B also suggests that the more patch-token, more directly RViT-based B-family design was stronger for four-class classification on this benchmark than the weaker A-family design.

6.2 What the explainability results mean

The explainability results should be interpreted according to what each method can measure. In the guided models, Grad-CAM++ was computed from the PFD-modified feature map, so it was more sensitive to the effects of guidance than in the ablation variants. This helps explain why Hybrid B showed the most tumour-centred Grad-CAM++ results in the reported samples.

Attention Rollout was less reliable for this hybrid setting. It captures only transformer attention, but the final prediction also depends on the CNN branch and the late-fusion head, which rollout cannot explain. In Hybrid A, the coarse 7×7 token grid further limits spatial precision. This helps explain why Hybrid A did not show a clear XAI improvement over Ablation A in the reported cases: the weak guidance changed one pathway, but the overall decision remained largely controlled by the ungated CNN branch. The explainability findings should therefore be treated as qualitative case-based evidence, not as full-dataset proof of localisation quality, especially because no ground-truth tumour masks were available for quantitative overlap testing.

6.3 Errors and uncertainty

The remaining errors were concentrated mainly at tumour subtype boundaries, especially between glioma, meningioma, and pituitary. This is consistent with a four-class 2D MRI task, where a single slice may not contain enough anatomical detail for reliable subtype separation.
The pituitary-as-meningioma pattern is also clinically plausible, because sellar and suprasellar meningiomas (tumours in or just above the pituitary area), including tuberculum sellae lesions at the skull base (tumours near the base of the brain, close to the pituitary area), can resemble pituitary adenomas on MRI and may be difficult to distinguish before surgery (Taylor et al., 1992; Kwancharoen et al., 2014).

The persistence of this confusion across all four variants suggests a limit of the input setting rather than a failure of one specific guidance module. Guidance can help reduce background-driven shortcut behaviour, but it cannot fully resolve genuine visual overlap when similar subtypes are represented by single 2D slices. This also helps explain why the stronger B guidance could improve tumour-centred localisation without removing subtype errors.

Hybrid B, and to a lesser extent Hybrid A, also produced some no tumour false positives. A cautious explanation is that MRI can contain suspicious non-neoplastic regions (areas that are not true tumours) or cystic changes (fluid-filled changes) that can look tumour-like on scans, so stronger guidance may sometimes treat them as tumour. Because this study used image-level labels rather than tissue confirmation for each highlighted region, the exact cause cannot be confirmed (Giraldo et al., 2023; Bonneville, Jäger and Smirniotopoulos, 2024; Pereira et al., 2021).

The uncertainty results add an important qualification. The pituitary example remained a wrong-but-confident case across all four variants, suggesting that the models had learned features strong enough to support a confident tumour decision, but not specific enough to separate the correct subtype. The uncertainty findings also did not show a clear advantage for the guided models: in this reported pituitary-to-meningioma error case, Hybrid B (0.9869) remained more confident when wrong than Ablation B (0.9632), while Hybrid A (0.9692) was also slightly more confident than Ablation A (0.9592). Uncertainty was therefore more informative when interpreted alongside the heatmaps than on its own.

6.4 Comparison with existing benchmark studies and research gap

Existing brain-tumour MRI studies on the Kaggle benchmark report strong classification accuracy with CNN, transformer, and hybrid architectures. However, these works differ in task definition, preprocessing, split construction, optimisation, checkpoint selection, and the level of implementation detail provided. The comparison in Table 6.1 is therefore used as a contextual benchmark rather than as a direct reproduction claim.

Table 6.1: Contextual comparison with existing benchmark studies

Study	Main model/task	Reported result	Limitation relative to this project	Relation to this project
Sarada et al. (2025)	Modified ResNet50V2 for brain-tumour MRI classification	96.33% accuracy	The study motivates the CNN branch, but the exact modified ResNet50V2 could not be reproduced because all layer-level and training details required for exact reconstruction were not explicitly specified. Shortcut-learning mitigation and branch-specific evidence analysis were also not central.	This project therefore used a reproducible timm ResNet50V2 backbone as the CNN feature extractor and integrated it into a matched hybrid and ablation framework. All four project variants achieved higher held-out accuracy, from 98.52% to 99.22%.
Al Bataineh et al. (2024)	Swin Transformer + ResNet50V2 hybrid on four-class brain MRI data	96.8% accuracy	The work supports CNN-transformer fusion, but the evaluation is mainly accuracy-centred and does not explicitly test whether the model relies on tumour-centred evidence rather than shortcut cues.	All four project variants achieved higher held-out accuracy under a fixed split and shared training protocol, while also adding Grad-CAM++, Attention Rollout, uncertainty analysis, and matched ablations.
Karagoz et al. (2024)	ResViT-based self-supervised hybrid CNN-transformer model	98.47% accuracy on the Kaggle brain tumour dataset	The method gives a strong hybrid benchmark, but its main emphasis is classification under limited-data conditions rather than explicit shortcut-learning analysis with guided and unguided matched variants.	All four project variants were numerically above the reported Kaggle accuracy, although the comparison remains contextual because the exact training and split conditions are not identical.
Krishnan et al. (2024)	Rotation-Invariant Vision Transformer (RViT), glioma against no tumour	98.6% accuracy	The paper gives the closest transformer-side architectural reference and provides the main RViT design details followed in this project. However, the reported experiment is binary glioma-against-no tumour classification, not the four-class subtype problem studied here.	The RViT branch was implemented following the rotation-aware design and reported hyperparameter details, but the architecture had to be revised to support ResNet50V2 fusion, PFD, GSTE, and the four-class hybrid setting. Hybrid A, Ablation A, and Ablation B were numerically above 98.6%; Hybrid B was slightly lower at 98.52%, so this is not a direct task-equivalent comparison.

This project	ResNet50V2-RViT hybrid variants with PFD-GSTE guidance and ablations	98.52% to 99.22% held-out accuracy; 98.49% to 99.20% macro-F1	The study is limited to one curated Kaggle benchmark and does not reproduce every previous model under identical external conditions.	The contribution is not only higher benchmark accuracy, but a reliability-focused evaluation of whether guided hybrid models can maintain strong performance while producing more tumour-centred evidence.

The comparison shows that the proposed variants are competitive with the reviewed CNN, transformer, and hybrid studies, and numerically exceed the main four-class benchmark accuracies reported in Chapter 2. This result should not be interpreted as a definitive state-of-the-art claim, because several prior studies do not provide enough implementation detail to reproduce their exact preprocessing, split policy, optimiser settings, augmentation, and checkpoint-selection procedure. In particular, Sarada et al. (2025) motivated the ResNet50V2 branch, but the exact modified network could not be reconstructed from the published description alone. A reproducible timm ResNet50V2 backbone was therefore used instead. Krishnan et al. (2024) provided the closest transformer reference, and the RViT branch in this project followed its rotation-aware design where applicable. The design was nevertheless revised because this project does not use RViT as a standalone binary classifier; it integrates RViT with ResNet50V2, PFD, GSTE, late fusion, and matched ablations for four-class classification.

The more important distinction is methodological. Prior studies mainly optimise and report classification performance. In contrast, this project evaluates whether strong performance is accompanied by tumour-centred evidence. This matters because high accuracy alone does not rule out shortcut learning: a model may classify the benchmark correctly while relying on background structure, acquisition artefacts, or other non-pathology cues. The proposed PFD-GSTE variants were introduced to test this specific reliability question, not simply to add architectural complexity.

The results show a separation between classification strength and evidence quality. Ablation B achieved the best held-out accuracy and macro-F1, indicating that the B-family patch-token design is a strong classifier even without explicit PFD-GSTE guidance. Hybrid B, however, produced the most tumour-centred Grad-CAM+ + behaviour in the reported qualitative cases, indicating that stronger guidance changed the evidence pathway more clearly than it improved the final score. This supports the main finding of the project: PFD-GSTE guidance can improve tumour-focused evidence in selected cases, but it does not guarantee a uniform gain in accuracy or uncertainty behaviour.

Overall, the comparison shows that the proposed ResNet50V2-RViT family is stronger than the reviewed four-class benchmark studies in this specific evaluation setting. The advantage is not only numerical. The project combines high held-out accuracy with ablations, branch-specific explainability, and uncertainty analysis, whereas the reviewed work mainly reports classification performance. Therefore, the contribution is best understood as a stronger benchmark-level study for this dataset: it does not claim clinical generalisation beyond the Kaggle setting, but it provides a more complete evaluation of performance and shortcut-learning risk than accuracy-only comparison.

6.5 Limitations and overall evaluation

Several limitations should remain clear when interpreting these findings. This project used one curated Kaggle benchmark and did not include external validation, so performance under different scanners, hospitals, acquisition settings, or clinical workflows was not tested. The explainability analysis was qualitative because tumour masks were not available for quantitative localisation scoring. MC Dropout provided a practical uncertainty estimate, but formal calibration metrics were not separately measured.A further implementation limitation is that the B-family used drop_last=True, whereas the A-family did not. This followed the implemented family designs, but it means the cleanest comparisons of the guidance modules are the within-family comparisons: Hybrid A against Ablation A, and Hybrid B against Ablation B. Within these limits, the study gives a clear answer to the project aim. The B-family design was the strongest classifier on this benchmark, with Ablation B achieving the highest held-out accuracy and macro-F1. The guided Hybrid B did not outperform its ablation numerically, but it produced the most tumour-centred Grad-CAM++ behaviour in the reported cases. This shows that PFD-GSTE guidance affected the evidence pathway more clearly than the final classification score.

The overall finding is therefore strong but precise. The project outperformed the reviewed four-class benchmark studies in this fixed-split evaluation and added reliability evidence that was largely absent from those works. However, subtype confusion and wrong-but-confident errors remained, so the results support partial mitigation of shortcut learning rather than a complete solution.

7. Conclusion

This project investigated whether pathology-guided hybrid CNN-transformer models could maintain strong four-class brain tumour MRI classification while producing more tumour-focused evidence as an experimental approach to reducing shortcut learning. To address this, a reproducible benchmark was prepared from the Kaggle Brain *Tumor* MRI dataset using offline preprocessing, duplicate handling, and fixed training, validation, and test splits. Two hybrid model variants and their corresponding ablations were then implemented and compared: Hybrid A, Hybrid B, Ablation A, and Ablation B.

At the benchmark level defined in this study, the project aims, objectives, and practical targets were addressed and largely achieved, with partial support for the main guidance hypothesis. All four models were trained and evaluated under one shared procedure, explainability outputs were generated on fixed reported cases, uncertainty summaries were reported, and the performance target was exceeded because more than one variant achieved a test macro-F1 above 0.98, with Ablation B reaching 0.9920. All four models also achieved strong held-out performance, with test accuracy ranging from 0.9852 to 0.9922 and macro-F1 ranging from 0.9849 to 0.9920. Numerically, Ablation B was the strongest model. Hybrid A and Ablation A were almost identical in performance, indicating that the weaker guidance design had little effect on final classification accuracy, whereas Hybrid B produced the most tumour-centred Grad-CAM++ results in the reported cases despite not outperforming Ablation B numerically.

The research question can therefore be answered partially. On this benchmark, stronger guidance improved tumour-focused evidence more clearly than classification performance. The study therefore supports the view that high benchmark accuracy alone is not sufficient to judge model reliability, and that explanation, rather than uncertainty, provided the clearest support when investigating possible shortcut learning in brain tumour MRI classification.

8. Future Work

A clear next step is to test the models beyond this single benchmark. The current findings were obtained on one aggregated dataset with one held-out test set, so performance under different scanners, hospitals, and acquisition conditions remains uncertain. External validation on independent brain MRI datasets would provide a stronger test of whether the reported behaviour remains stable when the data distribution changes, especially because shortcut-prone behaviour may become more visible when evaluation data differ from the training distribution, including in out-of-distribution (OOD) cases (Geirhos et al., 2020; Varoquaux and Cheplygina, 2022; Rouzrokh et al., 2022).

A second priority is to address the limits of single-slice subtype classification. In this project, many remaining errors occurred at tumour subtype boundaries, especially between pituitary and meningioma. Future work should therefore examine multi-slice, multi-sequence, or three-dimensional MRI inputs, so that subtype decisions are based on richer anatomical context rather than one 2D slice alone.
The guidance design also needs refinement. Hybrid B gave the most tumour-centred Grad-CAM++ results, but it did not outperform Ablation B numerically and produced more no tumour false positives. This suggests that stronger guidance can improve evidence focus while also making the model more sensitive to suspicious-looking non-tumour structure. Future work should therefore test softer mask weighting, better control of guidance strength, and designs that preserve more surrounding context.

The explainability and uncertainty pipeline should also be strengthened. Attention Rollout was limited for this hybrid setting, so future work should use branch-aware or fusion-aware explanation methods and, where tumour masks are available, compare heatmaps quantitatively against annotated regions rather than relying mainly on visual inspection (Adebayo et al., 2018; Arun et al., 2021). Uncertainty analysis should also be extended with calibration-focused evaluation, because some wrong subtype predictions remained highly confident (Gal and Ghahramani, 2016; Rajaraman, Ganesan and Antani, 2022; Penso, Frenkel and Goldberger, 2024).

9. Project Management Review

This project required significant replanning, so the main management task was adapting scope while still delivering a complete and academically focused outcome. The original disaster-detection idea was dropped after early supervision because suitable data were not realistically available and the topic was less well aligned with the final-year project aims. Refocusing on brain tumour MRI classification improved feasibility, but the project then became more demanding as the emphasis narrowed further to shortcut learning, explainability, and uncertainty.

The main weakness in management was early estimation of time and workload. As shown in Figure 9.1, the first Gantt chart was too optimistic, especially for literature review, model development, debugging, explainability, and later-stage analysis. The revised plans in Figures 9.2(a)-(d) were more realistic and show clear improvement in schedule management, with more time allocated for preprocessing, duplicate checking, fixed split preparation, reproducible experiments, figure production, and report writing. This shows growth in planning, but also highlights that feasibility and workload should have been tested more critically at the start.

A clear strength was quality control. Files, checkpoints, logs, exported metrics, figures, and report drafts were kept organised, which supported traceability and made corrections easier. Regular supervisor meetings also improved project control. They helped identify weak decisions early, sharpen the research direction, and improve communication by encouraging more focused and evidence-based questions. Another important lesson was the value of iterative review. Chapter plans, figures, and report wording were revised after Dr Koay's feedback, which improved clarity, structure, and alignment with the project aims. Maintaining dated drafts, meeting notes, and updated Gantt charts strengthened documentation, communication, and overall control of progress.

Resources such as Kaggle, published datasets, and online literature were sufficient, but more buffer time should have been reserved for designing and testing different experimental approaches, as well as for debugging and interpretation. Even so, the project was managed effectively enough to recover from major changes. Stronger early scoping and more realistic scheduling would have improved the process further.

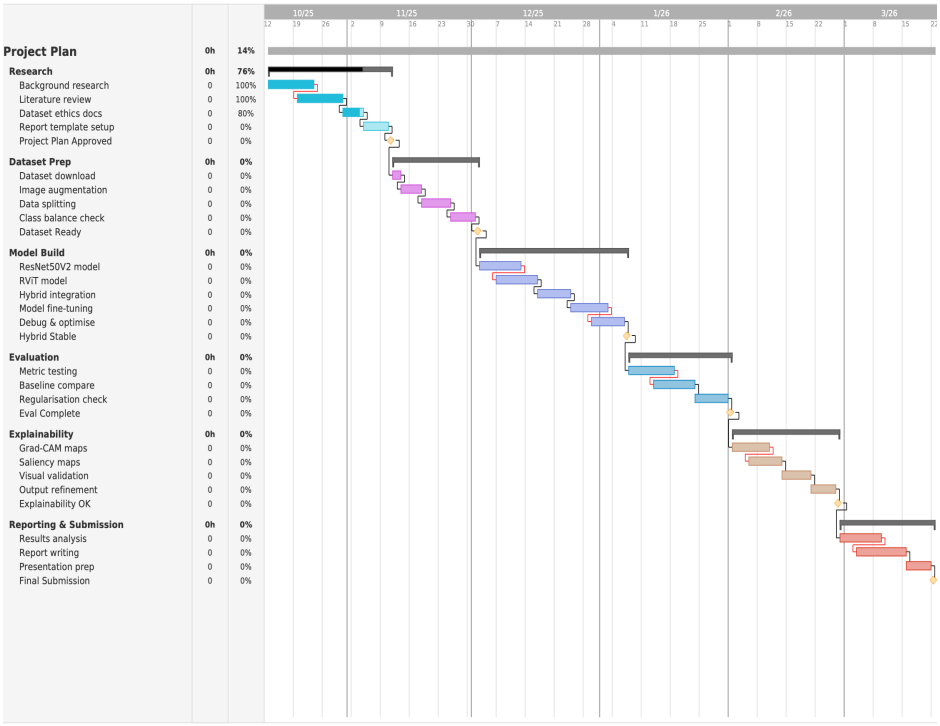
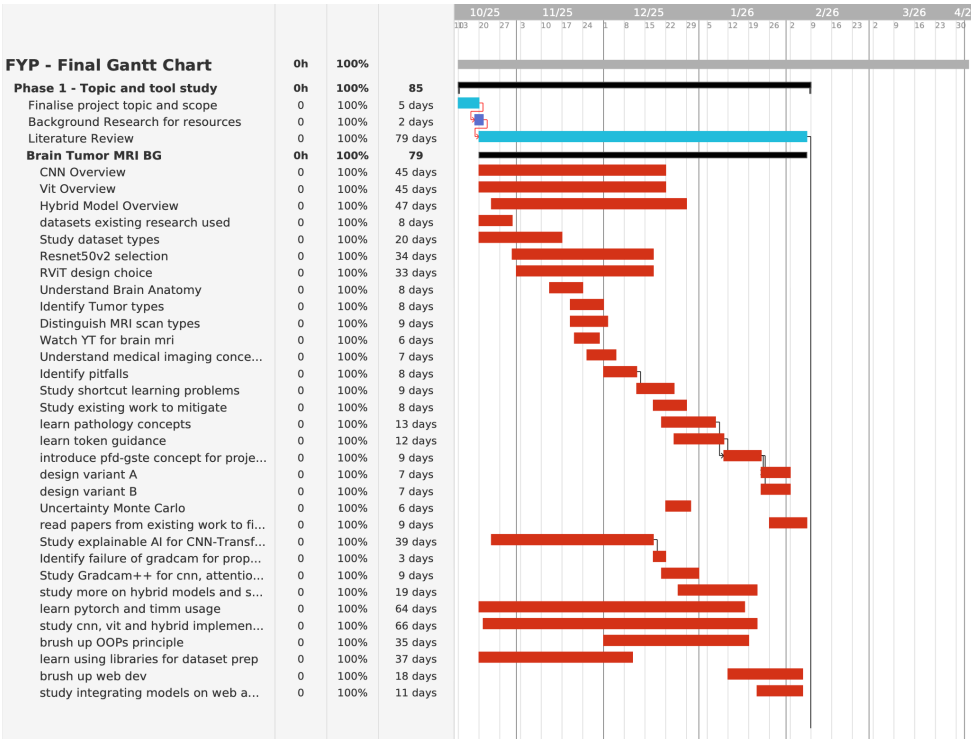
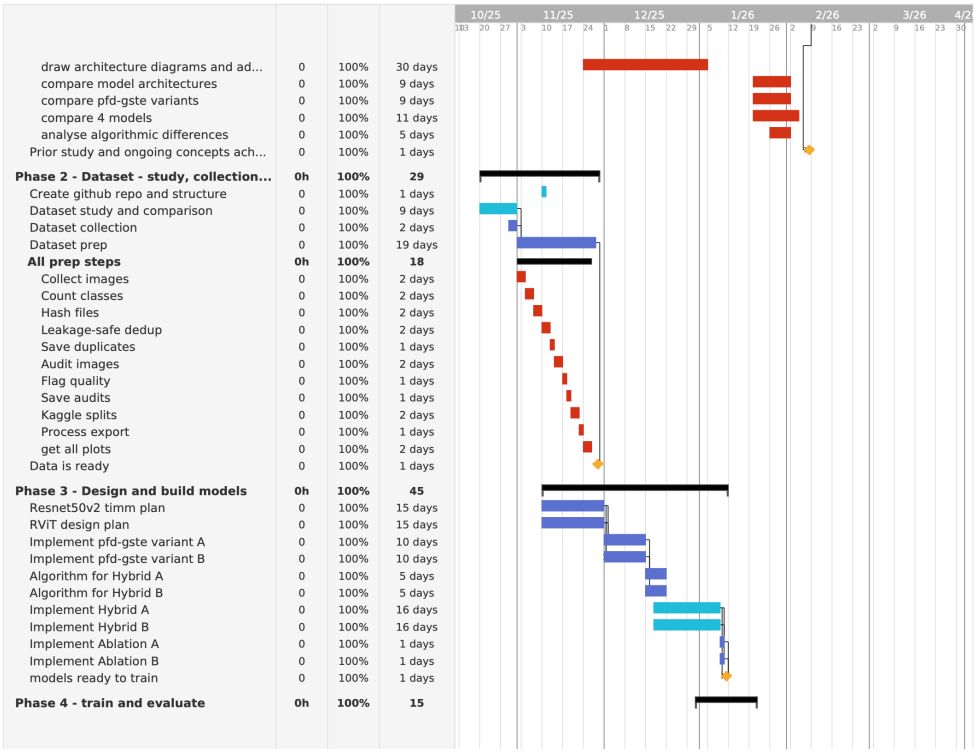


Figure 9.1: Initial project Gantt chart from the project outline.



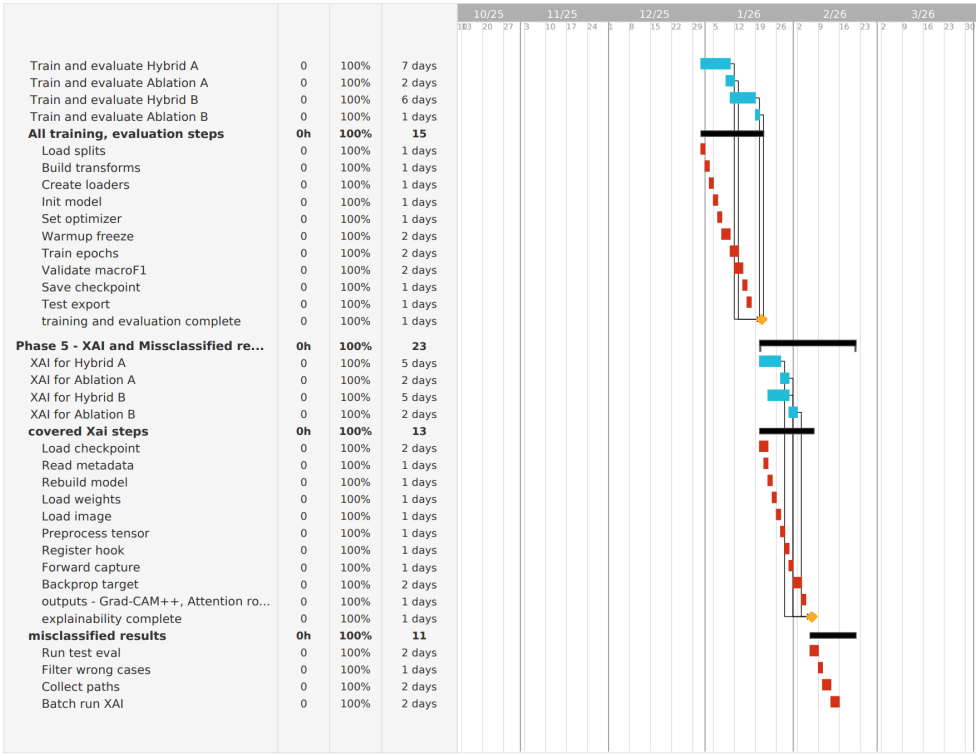
FYP - Final Gantt Chart	0h	100%	
Phase 1 - Topic and tool study	**0h**	**100%**	**85**
Finalise project topic and scope	0	100%	5 days
Background Research for resources	0	100%	2 days
Literature Review	0	100%	79 days
Brain Tumor MRI BG	**0h**	**100%**	**79**
CNN Overview	0	100%	45 days
Vit Overview	0	100%	45 days
Hybrid Model Overview	0	100%	47 days
datasets existing research used	0	100%	8 days
Study dataset types	0	100%	20 days
Resnet50v2 selection	0	100%	34 days
RViT design choice	0	100%	33 days
Understand Brain Anatomy	0	100%	8 days
Identify Tumor types	0	100%	8 days
Distinguish MRI scan types	0	100%	9 days
Watch YT for brain mri	0	100%	6 days
Understand medical imaging conce...	0	100%	7 days
Identify pitfalls	0	100%	8 days
Study shortcut learning problems	0	100%	9 days
Study existing work to mitigate	0	100%	8 days
learn pathology concepts	0	100%	13 days
learn token guidance	0	100%	12 days
introduce pfd-gste concept for proje...	0	100%	9 days
design variant A	0	100%	7 days
design variant B	0	100%	7 days
Uncertainty Monte Carlo	0	100%	6 days
read papers from existing work to fi...	0	100%	9 days
Study explainable AI for CNN-Transf...	0	100%	39 days
Identify failure of gradcam for prop...	0	100%	3 days
Study Gradcam++ for cnn, attentio...	0	100%	9 days
study more on hybrid models and s...	0	100%	19 days
learn pytorch and timm usage	0	100%	64 days
study cnn, vit and hybrid implemen...	0	100%	66 days
brush up OOPs principle	0	100%	35 days
learn using libraries for dataset prep	0	100%	37 days
brush up web dev	0	100%	18 days
study integrating models on web a...	0	100%	11 days

(a)

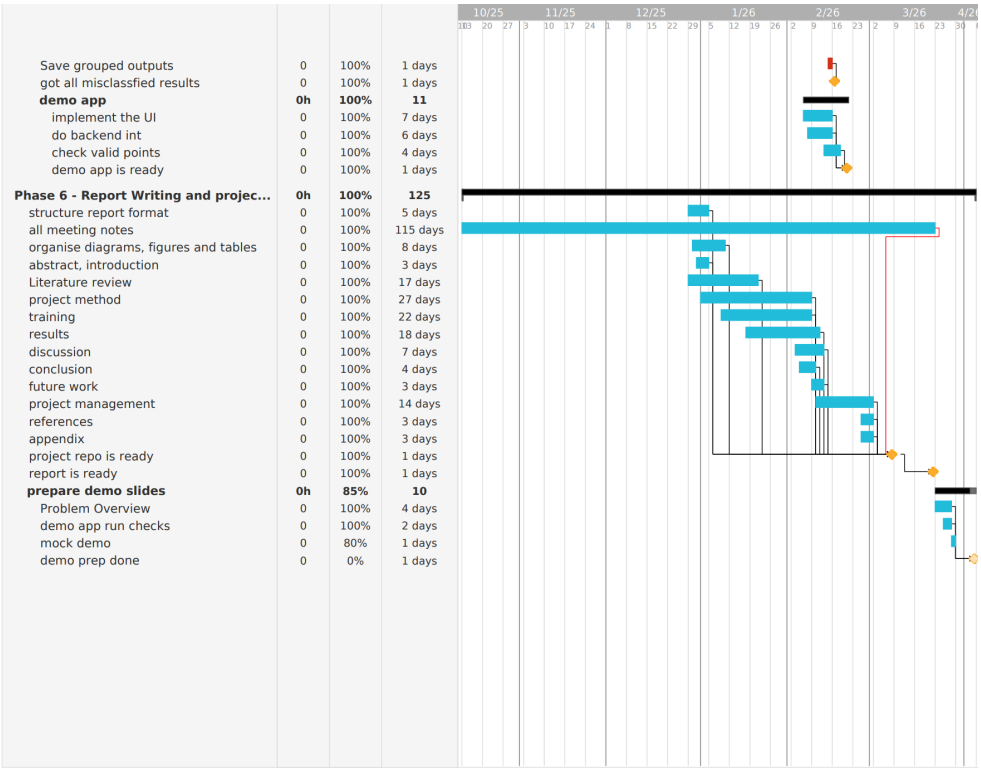


draw architecture diagrams and ad...	0	100%	30 days
compare model architectures	0	100%	9 days
compare pfd-gste variants	0	100%	9 days
compare 4 models	0	100%	11 days
analyse algorithmic differences	0	100%	5 days
Prior study and ongoing concepts ach...	0	100%	1 days
Phase 2 - Dataset - study, collection...	**0h**	**100%**	**29**
Create github repo and structure	0	100%	1 days
Dataset study and comparison	0	100%	9 days
Dataset collection	0	100%	2 days
Dataset prep	0	100%	19 days
All prep steps	**0h**	**100%**	**18**
Collect images	0	100%	2 days
Count classes	0	100%	2 days
Hash files	0	100%	2 days
Leakage-safe dedup	0	100%	2 days
Save duplicates	0	100%	1 days
Audit images	0	100%	2 days
Flag quality	0	100%	1 days
Save audits	0	100%	1 days
Kaggle splits	0	100%	2 days
Process export	0	100%	1 days
get all plots	0	100%	2 days
Data is ready	0	100%	1 days
Phase 3 - Design and build models	**0h**	**100%**	**45**
Resnet50v2 timm plan	0	100%	15 days
RViT design plan	0	100%	15 days
Implement pfd-gste variant A	0	100%	10 days
Implement pfd-gste variant B	0	100%	10 days
Algorithm for Hybrid A	0	100%	5 days
Algorithm for Hybrid B	0	100%	5 days
Implement Hybrid A	0	100%	16 days
Implement Hybrid B	0	100%	16 days
Implement Ablation A	0	100%	1 days
Implement Ablation B	0	100%	1 days
models ready to train	0	100%	1 days
Phase 4 - train and evaluate	**0h**	**100%**	**15**

(b)



(c)



(d)

Figure 9.2(a-d): Final Gantt charts showing the revised project schedule and phased implementation plan.

10. References

Abnar, S. and Zuidema, W. (2020) Quantifying attention flow in transformers, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190-4197. https://doi.org/10.18653/v1/2020.acl-main.385

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B. (2018) Sanity checks for saliency maps, *Advances in Neural Information Processing Systems*, *31*. https://doi.org/10.48550/arXiv.1810.03292

Al Bataineh, A.F., Nahar, K.M.O., Khafajeh, H., Samara, G., Alazaidah, R., Nasayreh, A., Bashkami, A., Gharaibeh, H. and Dawaghreh, W. (2024) Enhanced magnetic resonance imaging-based brain tumor classification with a hybrid Swin Transformer and ResNet50V2 model, *Applied Sciences*, *14*(22), 10154. https://doi.org/10.3390/app142210154

Amoo, M., Henry, J., Farrell, M. and Javadpour, M. (2023) Meningioma in the elderly, *Neuro-Oncology Advances*, *5*(Suppl 1), pp. i13-i25. https://doi.org/10.1093/noajnl/vdac107

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D. and Kalpathy-Cramer, J. (2021) Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging, *Radiology: Artificial Intelligence*, *3*(6), e200267. https://doi.org/10.1148/ryai.2021200267

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K. and Davatzikos, C. (2017) Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, *Scientific Data*, *4*, 170117. https://doi.org/10.1038/sdata.2017.117

Bolya, D., Fu, C., Dai, X., Zhang, P., Feichtenhofer, C. and Hoffman, J. (2023) Token merging: your ViT but faster, in *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2210.09461

Bonneville, F., Jäger, H.R. and Smirniotopoulos, J.G. (2024) Differential diagnosis of intracranial masses, in Hodler, J., Kubik-Huch, R.A. and Roos, J.E. (eds.) *Diseases of the Brain, Head and Neck, Spine 2024-2027: Diagnostic Imaging*. Cham: Springer, pp. 113-127. https://doi.org/10.1007/978-3-031-50675-8_8

Chattopadhyay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N. (2018) Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839-847. https://doi.org/10.1109/WACV.2018.00097

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Q. and Feng, Q. (2015) Enhanced performance of brain tumor classification via tumor region augmentation and partition, *PLOS ONE*, *10*(10), e0140381. https://doi.org/10.1371/journal.pone.0140381

Clark, A. and contributors (2024) *Pillow* [computer software], version 10.3.0. Available at: https://pillow.readthedocs.io/en/stable/ (Accessed: 13 November 2025).

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. and Prior, F. (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, *Journal of Digital Imaging*, *26*(6), pp. 1045-1057. https://doi.org/10.1007/s10278-013-9622-7

Comelli, I., Lippi, G., Campana, V., Servadei, F. and Cervellin, G. (2017) Clinical presentation and epidemiology of brain tumors firstly diagnosed in adults in the Emergency Department: a 10-year, single-center retrospective study, *Annals of Translational Medicine*, *5*(13), p. 269. https://doi.org/10.21037/atm.2017.06.12

da Costa-Luis, C.O. (2019) tqdm: a fast, extensible progress meter for Python and CLI, *Journal of Open Source Software*, *4*(37), 1277. https://doi.org/10.21105/joss.01277

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021) An image is worth 16×16 words: transformers for image recognition at scale, in *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. https://doi.org/10.48550/arXiv.2010.11929

European Union (2016) *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)*, *Official Journal of the European Union*, *L119*, pp. 1-88. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed: 23 November 2025).

Farhan, A.S., Khalid, M. and Manzoor, U. (2025) XAI-MRI: an ensemble dual-modality approach for 3D brain tumor segmentation using magnetic resonance imaging, *Frontiers in Artificial Intelligence*, *8*, 1525240. https://doi.org/10.3389/frai.2025.1525240

Gal, Y. and Ghahramani, Z. (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, PMLR, *48*, pp. 1050-1059. https://doi.org/10.48550/arXiv.1506.02142

Garcia-Atutxa, I., Martínez-Más, J., Bueno-Crespo, A. and Villanueva-Flores, F. (2025) Early-fusion hybrid CNN-transformer models for multiclass ovarian tumor ultrasound classification, *Frontiers in Artificial Intelligence*, *8*, 1679310. https://doi.org/10.3389/frai.2025.1679310

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F.A. (2020) Shortcut learning in deep neural networks, *Nature Machine Intelligence*, *2*(11), pp. 665-673. https://doi.org/10.1038/s42256-020-00257-z

Gholipour, A., Kehtarnavaz, N., Briggs, R., Devous, M. and Gopinath, K. (2007) Brain functional localization: a survey of image registration techniques, *IEEE Transactions on Medical Imaging*, *26*(4), pp. 427-451. https://doi.org/10.1109/TMI.2007.892508

Giraldo, G.S.P., Singer, L., Cao, T., Davis, K., Burroughs, C.J., Guo, E., Baca, R., Joseph, P., Chen, R., Kim, M., Aggarwal, N., Cohen, B.A. and Greenlee, E. (2023) Differential diagnosis of tumor-like brain lesions, *Neurology: Clinical Practice*, *13*(5), e200182. https://doi.org/10.1212/CPJ.0000000000200182

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. and Oliphant, T.E. (2020) Array programming with NumPy, *Nature*, *585*, pp. 357-362. https://doi.org/10.1038/s41586-020-2649-2

He, K., Zhang, X., Ren, S. and Sun, J. (2016) Identity mappings in deep residual networks, in Leibe, B., Matas, J., Sebe, N. and Welling, M. (eds.) *Computer Vision – ECCV 2016*. Cham: Springer, pp. 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

Ho, K.K.Y., Kaiser, U.B., Chanson, P., Gadelha, M., Wass, J., Nieman, L., Little, A., Aghi, M.K., Raetzman, L., Post, K., Raverot, G., Borowsky, A.D., Erickson, D., Castaño, J.P., Laws, E.R., Zatelli, M.C., Sisco, J., Esserman, L., Yuen, K.C.J., Reincke, M. and Melmed, S. (2023) Pituitary adenoma or neuroendocrine tumour: the need for an integrated prognostic classification, *Nature Reviews Endocrinology*, *19*(11), pp. 671-678. https://doi.org/10.1038/s41574-023-00883-8

Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K.Q. (2017) Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261-2269. https://doi.org/10.1109/CVPR.2017.243

Hunter, J.D. (2007) Matplotlib: a 2D graphics environment, *Computing in Science & Engineering*, *9*(3), pp. 90-95. https://doi.org/10.1109/MCSE.2007.55

Information Commissioner's Office (2023) *Guide to the UK GDPR*. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/ (Accessed: 23 November 2025).

Jain, S. and Wallace, B.C. (2019) Attention is not explanation, in *Proceedings of NAACL-HLT 2019*, pp. 3543-3556. https://doi.org/10.18653/v1/N19-1357

Karagoz, M.A., Nalbantoglu, O.U. and Fox, G.C. (2024) Residual Vision Transformer (ResViT) based self-supervised learning model for brain tumour classification, *arXiv* [Preprint]. https://doi.org/10.48550/arXiv.2411.12874

Khairandish, M.O., Gupta, R. and Sharma, M. (2020) A hybrid model of Faster R-CNN and SVM for tumor detection and classification of MRI brain images, *International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)*, *10*(3), pp. 6863-6876. https://doi.org/10.24247/ijmperdjun2020651

Koçak, B. (2022) Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics, *Diagnostic and Interventional Radiology*, *28*(5), pp. 450-462. https://doi.org/10.5152/dir.2022.211297

Krishnan, P.T., Krishnadoss, P., Khandelwal, M., Gupta, D., Nihaal, A. and Kumar, T.S. (2024) Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer, *Frontiers in Neuroinformatics*, *18*, 1414925. https://doi.org/10.3389/fninf.2024.1414925

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

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning, *Nature*, *521*(7553), pp. 436-444. https://doi.org/10.1038/nature14539

Lin, M., Weng, N., Mikolaj, K., Bashir, Z., Svendsen, M.B.S., Tolsgaard, M.G., Christensen, A.N. and Feragen, A. (2024) Shortcut learning in medical image segmentation, in Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K. and Schnabel, J.A. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Cham: Springer, pp. 623-633. https://doi.org/10.1007/978-3-031-72111-3_59

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) Swin Transformer: hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P. and Ellison, D.W. (2016) The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary, *Acta Neuropathologica*, *131*(6), pp. 803-820. https://doi.org/10.1007/s00401-016-1545-1

Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., Soffietti, R., von Deimling, A. and Ellison, D.W. (2021) The 2021 WHO Classification of Tumors of the Central Nervous System: a summary, *Neuro-Oncology*, *23*(8), pp. 1231-1251. https://doi.org/10.1093/neuonc/noab106

Maggio, I., Franceschi, E., Di Nunno, V. and Brandes, A.A. (2021) Meningioma: not always a benign tumor. A review of advances in the treatment of meningiomas, *CNS Oncology*, *10*(2), CNS72. https://doi.org/10.2217/cns-2021-0003

McKinney, W. (2010) Data structures for statistical computing in Python, in van der Walt, S. and Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*, pp. 56-61. https://doi.org/10.25080/Majora-92bf1922-00a

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., García, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G. and Wu, H. (2018) Mixed precision training, in *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. https://doi.org/10.48550/arXiv.1710.03740

Mongan, J., Moy, L. and Kahn, C.E. (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers, *Radiology: Artificial Intelligence*, *2*(2), e200029. https://doi.org/10.1148/ryai.2020200029

National Institute of Standards and Technology (2015) *Secure Hash Standard (SHS)*. FIPS PUB 180-4. Available at: https://csrc.nist.gov/publications/detail/fips/180/4/final (Accessed: 16 November 2025).

Nickparvar, M. (2021) *Brain Tumor MRI Dataset* [dataset]. Kaggle. Available at: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (Accessed: 13 October 2025).

Pallets (2024) *Flask documentation*. Available at: https://flask.palletsprojects.com/ (Accessed: 12 February 2026).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019) PyTorch: an imperative style, high-performance deep learning library, in Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, *32*, pp. 8024-8035. https://doi.org/10.48550/arXiv.1912.01703

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011) Scikit-learn: machine learning in Python, *Journal of Machine Learning Research*, *12*, pp. 2825-2830. Available at: https://jmlr.org/papers/v12/pedregosa11a.html (Accessed: 16 November 2025).

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

Pereira, R.G., Ribeiro, B.N.F., Hollanda, R.T.L., Almeida, L.B., Simeão, T.B. and Marchiori, E. (2021) Non-neoplastic intracranial cystic lesions: not everything is an arachnoid cyst, *Radiologia Brasileira*, *54*(1), pp. 49-55. https://doi.org/10.1590/0100-3984.2019.0144

Pollack, I.F., Agnihotri, S. and Broniscer, A. (2019) Childhood brain tumors: current management, biological insights, and future directions, *Journal of Neurosurgery: Pediatrics*, *23*(3), pp. 261-273. https://doi.org/10.3171/2018.10.PEDS18377

PyTorch (2024) *torchvision documentation*. Available at: https://docs.pytorch.org/vision/main/index.html (Accessed: 22 October 2025).

Rajaraman, S., Ganesan, P. and Antani, S. (2022) Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks, *PLOS ONE*, *17*(1), e0262838. https://doi.org/10.1371/journal.pone.0262838

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J. and Hsieh, C.-J. (2021) DynamicViT: efficient vision transformers with dynamic token sparsification, *Advances in Neural Information Processing Systems*, *34*. https://doi.org/10.48550/arXiv.2106.02034

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A., Etmann, C., McCague, C., Beer, L., Weir-McCall, J., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J., Sala, E. and Schönlieb, C.-B. (2021) 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*, pp. 199-217. https://doi.org/10.1038/s42256-021-00307-0

Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Vafaei Sadr, D., Golbabaee, M., Sreekumar, Y., Zagouras, A., Choy, G. and Erickson, B.J. (2022) Mitigating bias in radiology machine learning: 1. Data handling, *Radiology: Artificial Intelligence*, *4*(5), e210290. https://doi.org/10.1148/ryai.210290

Sánchez-Moreno, L., Perez-Peña, A., Duran-Lopez, L. and Dominguez-Morales, J.P. (2025) Ensemble-based convolutional neural networks for brain tumor classification in MRI: enhancing accuracy and interpretability using explainable AI, *Computers in Biology and Medicine*, *195*, 110555. https://doi.org/10.1016/j.compbiomed.2025.110555

Sarada, B., Reddy, K.N., Muktisingh, R., Babu, R. and Babu, B.S.S.V.R. (2025) Brain tumor classification using modified ResNet50V2 deep learning model, *International Journal of Computing and Digital Systems*, *17*(1), pp. 1-11. https://doi.org/10.12785/ijcds/1571021750

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626. https://doi.org/10.1109/ICCV.2017.74

Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and Summers, R.M. (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, *IEEE Transactions on Medical Imaging*, *35*(5), pp. 1285-1298. https://doi.org/10.1109/TMI.2016.2528162

Simonyan, K. and Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition, in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going deeper with convolutions, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B. and Liang, J. (2016) Convolutional neural networks for medical image analysis: full training or fine tuning?, *IEEE Transactions on Medical Imaging*, *35*(5), pp. 1299-1312. https://doi.org/10.1109/TMI.2016.2535302

Tan, M. and Le, Q. (2019) EfficientNet: rethinking model scaling for convolutional neural networks, in *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, PMLR, *97*, pp. 6105-6114. https://doi.org/10.48550/arXiv.1905.11946

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

Tian, Y. (2025) Trade-off analysis of classical machine learning and deep learning models for robust brain tumor detection: benchmark study, *JMIR AI*, *4*, e76344. https://doi.org/10.2196/76344

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H. (2021) Training data-efficient image transformers & distillation through attention, in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, PMLR, *139*, pp. 10347-10357. https://doi.org/10.48550/arXiv.2012.12877

University of Hertfordshire (2021) *Data Protection Policy and Privacy Statement*. Available at: https://www.herts.ac.uk/__data/assets/pdf_file/0017/233090/IM08-Data-Protection.pdf (Accessed: 24 November 2025).

Varoquaux, G. and Cheplygina, V. (2022) Machine learning for medical imaging: methodological failures and recommendations for the future, *npj Digital Medicine*, *5*, 48. https://doi.org/10.1038/s41746-022-00592-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need, *Advances in Neural Information Processing Systems*, *30*, pp. 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

Villanueva-Meyer, J.E., Mabray, M.C. and Cha, S. (2017) Current clinical brain tumor imaging, *Neurosurgery*, *81*(3), pp. 397-415. https://doi.org/10.1093/neuros/nyx103

Weller, M., van den Bent, M., Preusser, M., Le Rhun, E., Tonn, J.C., Minniti, G., Bendszus, M., Balana, C., Chinot, O., Dirven, L., French, P., Hegi, M.E., Jakola, A.S., Platten, M., Roth, P., Rudà, R., Short, S., Smits, M., Taphoorn, M.J.B., von Deimling, A., Westphal, M., Soffietti, R., Reifenberger, G. and Wick, W. (2021) EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood, *Nature Reviews Clinical Oncology*, *18*(3), pp. 170-186. https://doi.org/10.1038/s41571-020-00447-z

Wightman, R. (2019) *timm: PyTorch image models* [computer software]. Available at: https://github.com/rwightman/pytorch-image-models (Accessed: 18 December 2025).

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. and Zhang, L. (2021) CvT: introducing convolutions to vision transformers, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22-31. https://doi.org/10.1109/ICCV48922.2021.00009

Xia, T., Chartsias, A. and Tsaftaris, S.A. (2020) Pseudo-healthy synthesis with pathology disentanglement and adversarial learning, *Medical Image Analysis*, *64*, 101719. https://doi.org/10.1016/j.media.2020.101719

Yadav, R.K., Kumar, M. and Nandi, A. (2025) 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. https://doi.org/10.1007/s44196-025-00981-7

Yazdan, S.A., Ahmad, R., Iqbal, N., Rizwan, A., Khan, A.N. and Kim, D.-H. (2022) An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD, *Tomography*, *8*(4), pp. 1905-1927. https://doi.org/10.3390/tomography8040161

11. Bibliography

Beaber, B. (2020) *MRI brain anatomy explained by neurologist* [YouTube video]. Available at: https://www.youtube.com/watch?v=Lmt1q9k78AM (Accessed: 17 November 2025).

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning*. Cambridge, MA: MIT Press. ISBN 9780262035613.

Hugging Face (2019) *timm/resnetv2_50x1_bit.goog_in21k_ft_in1k* [model weights]. Available at: https://huggingface.co/timm/resnetv2_50x1_bit.goog_in21k_ft_in1k (Accessed: 14 February 2026).

Kleinberg, J. and Tardos, É. (2006) *Algorithm design*. 1st edn. Boston, MA: Pearson Education/Addison-Wesley. ISBN 0-321-29535-8.

Leicester Medical School Radiology (2021) *Introduction to MRI of the brain* [YouTube video]. Available at: https://www.youtube.com/watch?v=Co7Qyu21QKU (Accessed: 22 November 2025).

Zhang, A., Lipton, Z.C., Li, M. and Smola, A.J. (2023) *Dive into deep learning*. Cambridge: Cambridge University Press. ISBN 9781009389433

12. Appendices

A1. Training and model hyperparameters for Hybrid models and their ablations

Table A1.1: Hybrid A and B

Category	Hyperparameter	Hybrid A	Hybrid B
Platform	Compute environment	Kaggle, single GPU	Kaggle, single GPU
Platform	GPU used	NVIDIA Tesla P100	NVIDIA Tesla P100
Reproducibility	Random seed	42	42
Training length	Max epochs	100	100
Training batch	Batch size	32	32
Data loading	Workers / pinned memory	2 workers, pin_memory=True	2 workers, pin_memory=True
Data loading	Train shuffle	True	True
Data loading	drop_last (train loader)	False	True
Inputs	Image format	224×224 RGB (preprocessed offline)	224×224 RGB (preprocessed offline)
Augmentation (train only)	Rotation	±15°	±15°
Augmentation (train only)	Horizontal flip	p = 0.5	p = 0.5
Augmentation (train only)	Translation	up to 5%	up to 5%
Augmentation (train only)	Gaussian noise	std=0.02, p=0.5	std=0.02, p=0.5
Normalisation	Mean / Std	(0.5, 0.5, 0.5) / (0.5, 0.5, 0.5)	Same
Loss	Training loss	Cross-entropy, label_smoothing=0.05	Same
Loss	Val/test loss reporting	Plain cross-entropy (no smoothing)	Same
Optimiser	Type	AdamW	AdamW
Optimiser	CNN learning rate	1×10^{-4}	1×10^{-4}
Optimiser	Transformer/fusion learning rate	5×10^{-4}	5×10^{-4}
Regularisation	Weight decay	0.01 (bias/norm/1D excluded)	Same
Scheduler	LR schedule	Cosine annealing	Cosine annealing
Scheduler	Minimum LR	$\eta_{min} = 1\times10^{-6}$	$\eta_{min} = 1\times10^{-6}$
Fine-tuning	Warmup (CNN frozen)	5 epochs	5 epochs
Stability	Gradient clipping	max_norm = 1.0	max_norm = 1.0
Model backbone	CNN	ResNet50V2 (pretrained)	ResNet50V2 (pretrained)
Transformer config	embed_dim / depth / heads / mlp_dim	142 / 10 / 10 / 480	142 / 10 / 10 / 480
Dropout	attn_dropout / vit_dropout	0.1 / 0.1	0.1 / 0.1
Fusion	fusion_dim / fusion_dropout	256 / 0.5	256 / 0.5
Internal invariance	Rotations used inside model	(0, 1, 2, 3)	(0, 1, 2, 3)
Selection	Early stopping metric	Val macro-F1, patience=10	Val macro-F1, patience=10
Transformer input	Patch size	Not applicable	16×16

Table A1.2: Ablation A and B

Aspect	Ablation A	Ablation B
Based on	Hybrid A	Hybrid B
Main change	PFD-GSTE guidance removed	PFD-GSTE guidance removed
Purpose	Test Variant A without guidance	Test Variant B without guidance
Train loader drop_last	False (same as Hybrid A)	True (same as Hybrid B)
All other training settings	Same as Hybrid A in Table A1.1	Same as Hybrid B in Table A1.1

Apart from removing the guidance modules, the ablation models keep the same training length, batch size, seed, optimiser, scheduler, augmentation, normalisation, backbone settings, transformer settings, fusion settings, and early-stopping rule as their corresponding hybrid models.

AMP and CNN BatchNorm freezing were not used in any of the four reported model runs.

A2. External software libraries, frameworks and implementation tools

The external software libraries and frameworks used in this project are:

- PyTorch: used for tensor computation, model building, training, inference, dropout-based uncertainty estimation, and GPU execution (Paszke et al., 2019).
- torchvision: used for image transformations and computer-vision utilities in the training data pipeline (PyTorch, 2024).
- timm: used to instantiate the pretrained ResNet50V2 backbone employed in the hybrid architectures (Wightman, 2019).
- NumPy: used for numerical operations, array manipulation, statistics, and intermediate result handling (Harris et al., 2020).
- pandas: used to create and manage CSV split files, dataset records, preprocessing tables, and experiment summaries (McKinney, 2010).
- scikit-learn: used for stratified train-validation splitting and evaluation metrics such as classification reports and confusion matrices (Pedregosa et al., 2011).
- Matplotlib: to generate training curves, confusion matrices, dataset figures, and XAI overlay visualisations (Hunter, 2007).
- Pillow: used for image loading, EXIF-safe correction, cropping, resizing, and RGB conversion during preprocessing and inference (Clark and contributors, 2024).
- Flask: to build the supporting prototype web interface for qualitative inspection of uploaded MRI images and model outputs (Pallets, 2024).
- tqdm: to provide progress bars during longer preprocessing and related utility scripts
- (da Costa-Luis, 2019).
- The architecture diagrams presented in this report were drawn using diagrams.net and the Concepts application on iPad.

A3. Source code access and repository guide

The full project source code is stored in a private GitHub repository for academic assessment and supervision.

- Any marker or authorised reader who wishes to review the code should email rb23ack@herts.ac.uk with their GitHub username and request access.
- If the reader does not already have a GitHub account, they should first create an account at https://github.com/login, and then send their GitHub username by email to request collaborator access to the repository.
- After being added as a collaborator, the reader should accept the GitHub invitation and follow the setup instructions provided in the repository README.
- The repository includes the preprocessing pipeline, training scripts, explainability scripts, trained-model loading for the demo application, and related documentation.

A4. Reproducibility and setup guide

To reproduce the project, a reader should:

1. clone the repository and install the required dependencies using the instructions in the project README;
2. obtain the Brain *Tumor* MRI dataset from Kaggle; the dataset link is provided in the README, and the extracted files should be placed under data/raw/brain-tumor-mri-dataset/;
3. run scripts/dataset_prep.py to perform leakage-safe deduplication, tight-crop preprocessing, and CSV split generation; this creates the processed images under data/processed/tightcrop/ and the split files under data/splits/tightcrop/;
4. run the required training script for Hybrid A, Hybrid B, Ablation A, or Ablation B using the fixed CSV splits;
5. use the saved best checkpoint and exported outputs (best_model.pt, history.csv, curves, confusion matrix, and metrics.json) for evaluation;
6. optionally run the XAI scripts to generate Grad-CAM++ and Attention-Rollout outputs from the trained checkpoints;
7. run the Flask demo application from the webapp/ directory for qualitative inspection of the trained models.

The README provides the commands, dependency setup, operating-system notes, Git LFS (Large File Storage) instructions, and model-loading details required to complete these steps.