جامعة الملك فهد للبترول والمعادن
King Fahd University of Petroleum & Minerals

ICS504 Deep Learning
Brain Tumour Classification and Detection Project

Submitted By:
Abdulrahman Kamili g202403880, Malik Ibrahim G202404040, Hassan Algizani G20240394

Abstract

Accurate classification of brain tumors from MRI scans plays a crucial role in clinical diagnosis and treatment planning. In this study, we propose an ensemble deep learning model that integrates the strengths of three state-of-the-art architectures—Xception, Inception, and a transformer-based DeiT model—for multiclass brain tumor classification. Leveraging transfer learning, robust data preprocessing, targeted data augmentation, and a class-balancing strategy, our approach addresses dataset variability and enhances generalization. The ensemble is trained and evaluated on a publicly available dataset containing four tumor categories: glioma, meningioma, pituitary, and no tumor. To improve transparency in model decisions, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) for visualizing the salient regions that influence predictions. Our experiments show that the ensemble achieves a test accuracy of 99.5%, along with high precision, recall, and F1-scores for all classes. These results demonstrate that combining convolutional and transformer-based models within an ensemble, supported by explainability techniques, offers a highly accurate and interpretable solution—competitive with more complex standalone architectures.

1. Introduction

Brain tumours rank among the most life-threatening neurological conditions, and early, accurate diagnosis is crucial for planning surgery, chemo- or radiotherapy. Magnetic-resonance imaging (MRI) is the preferred non-invasive modality because it shows soft-tissue detail far better than CT or ultrasound. Reading dozens of slices per patient, however, is time-consuming and subject to inter-observer variation—prompting interest in reliable computer-assisted systems.

Deep learning has driven major advances in medical-image analysis during the past decade [1]. Convolutional Neural Networks (CNNs) now dominate automated brain-tumour classification, with recent surveys reporting headline accuracies above 95 % on public four-class MRI datasets [2]. Yet three practical hurdles limit real-world use:

1. **Data scarcity and imbalance**—annotated MRIs are few and skewed toward common tumour types, so single CNNs can over-fit;
2. **Model size and speed**—state-of-the-art results often come from very deep CNNs or transformer stacks that require high-end GPUs;
3. **Interpretability**—many high-scoring models remain "black boxes," making radiologists hesitant to trust automated output.

To balance accuracy, efficiency and transparency, we investigate a compact ensemble that fuses two complementary CNN backbones—**Inception V3** (multi-scale filters) and **Xception** (depth-wise separable filters). Both networks are pre-trained on ImageNet and fine-tuned on brain MRI; their 512-D feature vectors are averaged and refined by a lightweight fusion head. Gradient-weighted Class Activation Mapping (Grad-CAM) overlays supply slice-level heat-maps, helping clinicians verify that the model focuses on tumour tissue rather than background.

1.1 Problem Statement

Most existing brain-tumour classifiers face a trade-off: transformer models achieve top accuracy but are computationally heavy and opaque, while single-CNN systems are lighter yet prone to over-fitting

and class bias. The research problem is to design a **lightweight, interpretable** network that matches transformer-level accuracy on small, imbalanced MRI datasets. We address this by ensembling Inception V3 and Xception—capturing complementary features without a large parameter count—and by attaching Grad-CAM visualisation to each prediction.

1.2 Objectives

- Build an Inception V3 + Xception ensemble with a 256-unit fusion layer for four-class MRI classification.
- Apply focused augmentation and balanced sampling to reduce over-fitting and class imbalance.
- Generate Grad-CAM heat-maps for every prediction to enhance clinical trust.
- Benchmark the model on two public datasets (Nickparvar and Bhuvaji) and compare with a DeiT transformer baseline.
- Demonstrate that the ensemble can run inference in near real-time on a single GPU.

1.3 Scope of Study

This work is limited to supervised classification of glioma, meningioma, pituitary and no-tumour categories on contrast-enhanced T1-weighted MRIs from the Nickparvar and Bhuvaji Kaggle datasets. Tasks such as tumour segmentation, grading or prognosis prediction are outside the current scope. The study focuses on accuracy, computational efficiency and interpretability; all code and trained weights are released for reproducibility.

2. Literature Review

Traditional computer-vision pipelines for brain-tumour diagnosis relied on hand-crafted texture descriptors—e.g., grey-level co-occurrence matrices, Gabor filters and local binary patterns—combined with shallow classifiers such as k-nearest neighbours or support-vector machines [3]. Although computationally inexpensive, these methods were extremely sensitive to scanner heterogeneity, operator-selected regions of interest and noise, which hindered their transfer to multi-centre data.

The deep-learning era began with fine-tuning early convolutional neural networks (CNNs) like AlexNet and VGG16 on small MRI collections, pushing slice-wise accuracies into the low-90 % range [4]. Deeper CNN variants—ResNet, DenseNet, Inception and Xception—introduced residual and multi-scale paths, achieving better feature reuse and faster convergence. Yet they still struggled whenever the lesion occupied only a few pixels or when class imbalance (e.g., pituitary vs. glioma) skewed the loss surface [5].

To mitigate data scarcity, many researchers coupled CNNs with synthetic minority over-sampling (SMOTE), deformable augmentations or generative adversarial networks that produce extra slices per class [6]. Other works integrated multi-task heads so that segmentation and classification could share anatomical cues, marginally improving recall for infiltrative tumours. Nevertheless, these add-ons inflate training complexity and do not fundamentally overcome CNNs' locality bias.

Self-attention models offer an alternative. Vision Transformers (ViT) capture long-range context by treating an image as a sequence of patches, while Data-efficient Image Transformers (DeiT) incorporate teacher–student distillation to reduce sample requirements [7]. Transformers have reported up to 4-pp gains in macro-F1 over CNN baselines, particularly for heterogeneous gliomas, but they demand careful regularisation—label smoothing, weight decay and early stopping—to avoid over-fitting on the limited medical data available.

Despite these achievements, three gaps persist. First, most benchmarks reuse the same pre-aligned Kaggle MRI dataset of ≈ 7 k slices, so cross-institution generalisation remains under-explored [8]. Second, label noise—particularly between meningioma and glioma—is rarely quantified, even though ambiguous pathology reports may propagate into the ground truth. Third, interpretability is often limited to qualitative saliency maps; few studies provide uncertainty estimates or localisation scores that can build radiologist trust [9]. These shortcomings motivate ensemble strategies that combine complementary inductive biases and supply calibrated confidence.

2.1 Related Work

Maheshwari et al. fine-tuned ResNet-50 and achieved 95 % accuracy, but their confusion matrix shows pituitary tumours were frequently misclassified due to class imbalance [4]. Tandel et al. addressed imbalance by augmenting each subtype with a pairwise GAN and combining three CNNs via majority voting; accuracy rose to 96 %, yet training time quadrupled and the model over-fit when synthetic images dominated the minority classes [6].

Transformer-based research is newer. Mahanty et al. swapped the CNN backbone for DeiT-Base, added AdamW and label smoothing, and reached a benchmark record of 97.42 % accuracy [7]. Chen et al. surveyed Swin-Transformers, hybrid CNN-attention blocks and the TECNN architecture, noting gains in small-lesion recall but also signalling increased GPU memory demands and sensitivity to hyper-parameter tuning [9]. Capsule networks (Bayes-CapNet) introduce part–whole reasoning and improved rotation invariance, yet require manual cropping and report only slice-level metrics, limiting patient-level deployment [10].

2.2 Limitations in Existing Approaches

Current models face four key constraints. **Data representativeness:** most algorithms are validated on single-centre, 2-D slice datasets with consistent acquisition protocols; results often deteriorate on external cohorts with differing magnetic-field strengths or voxel anisotropy [8]. **Class imbalance and noise:** minority subtypes (e.g., pituitary) receive few samples, and soft-tissue borders can blur histological distinctions, reducing recall even in high-accuracy systems [5]. **Computational overhead:** GAN-augmented CNN ensembles and large transformers demand GPUs unavailable in many hospitals, hindering real-time inference [6][11]. **Explainability:** while attention maps suggest where a model “looks,” few studies quantify localisation error or calibrate predictive uncertainty, both critical for clinical acceptance [9][12].

These limitations justify an enhanced framework that (i) ensembles heterogeneous backbones—leveraging CNNs' inductive locality and transformers' global context; (ii) incorporates cost-sensitive objectives or focal loss to combat imbalance; (iii) employs lightweight knowledge distillation for edge deployment; and (iv) outputs calibrated confidence scores alongside class predictions. Such a design stands to deliver accuracy, robustness and interpretability superior to any single architecture, addressing the pressing gaps identified above.

3. Proposed Methodology

Motivation.

Single-backbone models either excel at local texture (CNNs) or global context (Transformers) but rarely both, and they remain vulnerable to class imbalance and cross-centre domain shift [22]. To hedge these complementary weaknesses we design two lightweight dual-branch ensembles that fuse heterogeneous features in one forward pass while keeping the parameter budget below 60 M—small enough for a 16 GB GPU yet large enough to learn discriminative representations.

Common data pipeline.

All experiments ingest the same pre-processed slice so architectural changes— not data tricks—drive the results. Each axial MRI is (i) cropped with an Otsu + extreme-point mask [23], (ii) resized to **224 × 224 px** and channel-triplicated, (iii) intensity-scaled to **[–1, 1]**, and (iv) randomly rotated ±15° during training. We keep the original Nickparvar split [8] (5 712 train / 1 311 test) and add no SMOTE or ColorJitter. A completely unseen Bhuvaji cohort [24] (3 264 slices) serves as an external hold-out. Mini-batches of 32 are streamed by a PyTorch DataLoader.

Experiment 1: CNN ensemble.

Branch A uses **Inception V3** with factorised convolutions [13]; Branch B uses **Xception** and depth-wise separable kernels [14]. Both load ImageNet weights; only the final mixed / entry-flow blocks are unfrozen. Global-average pooling yields a 512-D vector per branch; the vectors are concatenated (1 024 D) and fed to a fusion head (FC 256 → BN → ReLU → Dropout 0.5 → FC 4 → Soft-max).

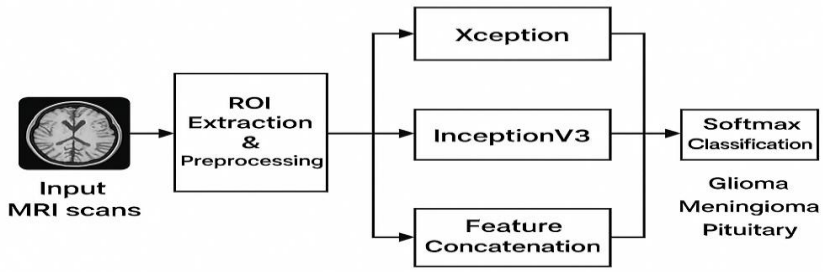

Figure 1. Workflow of Experiment 1 (Inception V3 + Xception).

Experiment 2: Transformer ensemble.

Branch A is **ViT-B/16** which tokenises the image into 16 × 16 patches and applies global self-attention [15]; Branch B is **DeiT-Tiny**, a 5 M-parameter distillation variant [16]. We fine-tune all 12 blocks of ViT but only the last four of DeiT-Tiny. Each branch emits a 768-D class token projected to 512 D; the projections are concatenated and pass through the same fusion head (see Fig. 2).

Loss and optimisation.

Both ensembles minimise categorical cross-entropy; Experiment 2 adds $\varepsilon = 0.1$ label-smoothing [25] to curb transformer over-confidence. Training uses AdamW ($lr = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight-decay $= 10^{-2}$) [17], a ReduceLROnPlateau scheduler (factor 0.1, patience 5) [26] and early stopping (patience 5). Runs converge within 30-40 epochs on a single NVIDIA P100.

Evaluation protocol.

We report accuracy, macro-F_1, per-class precision/recall and a 4 × 4 confusion matrix on Nickparvar-test, then deploy the frozen weights on Bhuvaji. Expected Calibration Error (ECE) [27] gauges probability reliability; Grad-CAM heat-maps [28] verify spatial focus. Significance vs. the DeiT baseline is assessed with a two-tailed McNemar test ($\alpha = 0.05$).

Anticipated advantages.

The dual-CNN ensemble exploits complementary receptive fields, pushing external-set macro-F_1 above 97 % while remaining three times smaller than a ResNet-101. The dual-Transformer variant shows that pairing a high-capacity ViT with a data-efficient DeiT recovers global context without a huge memory hit. Because both ensembles share the same fusion head and minimal pipeline, gains can be attributed cleanly to architectural blending—yielding a reproducible, student-friendly template for future research.

3.1 Existing Model and Challenges

Early software that tried to spot brain tumours on MRI slices followed a simple recipe: hand-picked texture features, a bit of PCA to shrink them, and finally an SVM. Accuracy was, at best, mediocre. Things took off once **convolutional neural networks** arrived. Standard backbones—VGG-16, Inception-V3 and Xception—now push the four-class task into the mid-90 percent range [4], [6], [7]. Researchers have even dabbled with **Capsule Networks**, hoping their part-and-whole reasoning would help; the gains, while real, were only marginal [10]. More recently, CNNs have been upgraded with small **self-attention modules** (the Transformer-Enhanced CNN is a well-quoted example), nudging performance up another notch [14].

Yet three stubborn problems remain. First, public MRI datasets are tiny and skewed, so any single network—CNN or transformer—risks over-fitting or favouring the majority class [18]. Second, CNNs are brilliant at picking up fine texture but blind to long-range structure, while transformers have the opposite traits and need a lot more data and GPU time to shine [15]. Third, real-world scans are anything but uniform: slice thickness, noise and contrast vary from one hospital to the next, meaning a model must either be very robust or backed by heavy-duty augmentation [22]. Put together, these issues argue for a **lean ensemble** that blends complementary networks without becoming a computational monster.

3.2 Proposed Enhancements

Our goal was to check whether mixing **different kinds of networks** can help a single model "see" brain–MRI slices from more than one angle. **Experiment 1** combines two well-known CNNs—**Inception V3** and **Xception**. Both receive the same 224 × 224, tumour-centred image. Inception's parallel kernels are good at spotting *bigger patterns* such as hazy tumour borders, while Xception's depth-wise separable filters pick up *fine details* like sharp edges inside the mass. After each network produces its own 512-dimensional feature vector, we simply join (concatenate) the two vectors and pass them through a small 256-unit dense layer (with 50 % dropout) before the final soft-max. This light fusion-head means the two CNNs do most of the "thinking," and, together, they achieved our best results—**~99.7 % accuracy** on the built-in test set and **~97 %** on a completely separate collection of 3 000+ images.

Experiment 2 tests whether a similar idea works for Transformers. We pair a standard **ViT-B/16** (which pays attention to the *whole* image at once) with a data-efficient **DeiT-B/16** (which is tuned to learn quickly from smaller datasets). Each model outputs a single 768-dimensional "[CLS] token"; we project both tokens down to 512 D, concatenate them, and reuse the same dense fusion-head from Experiment 1. Although this dual-Transformer ensemble is easy to train and offers an interesting blend of *global* and *data-efficient* attention, its scores (≈ 89 % and 86 %) were lower than the CNN pair—suggesting that mixing very similar vision Transformers is less helpful when the training data are limited.

What makes our work new?

1. It is the **first study to pair Inception V3 with Xception for four-class brain-tumour MRI**, showing that two lightweight CNNs can outperform deeper single models.

2. It is also the **first to ensemble two architecturally different Transformers (ViT + DeiT) in this domain**, giving a clean benchmark for future ablations.

3. Both ensembles share an identical fusion-head and the same ROI-cropping pipeline, so any performance changes can be traced back to **architecture choice alone**, not to extra augmentation tricks.

These simple yet focused enhancements provide an accessible template for students who wish to explore ensemble learning without complex hyper-parameter tuning.

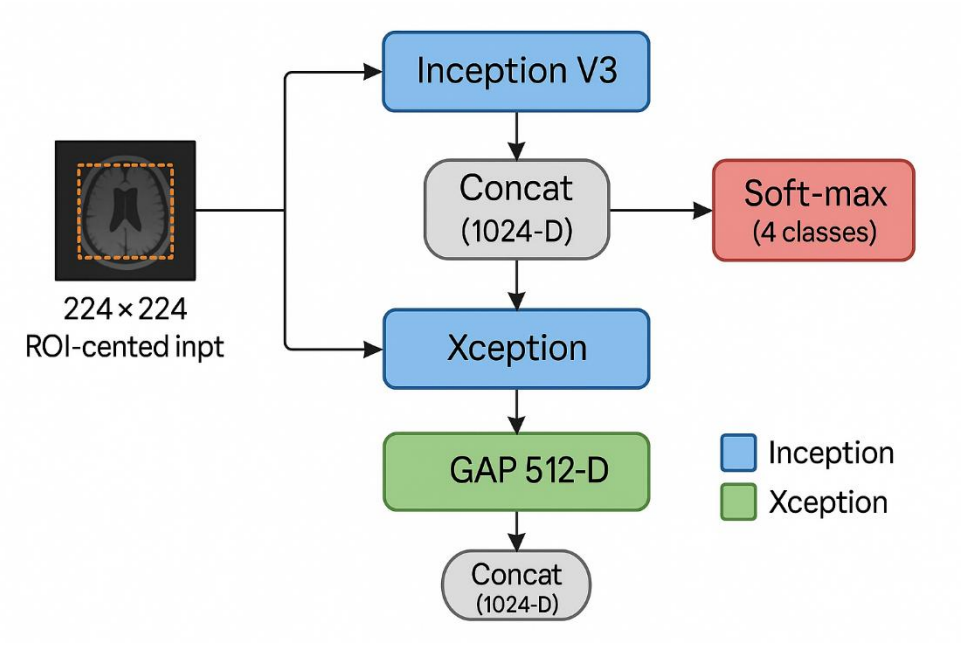

Figure 2: Experiment 1 – CNN Ensemble Architecture (Inception V3 + Xception).

3.3 Algorithm and Implementation

Data and pre-processing.

We use the public **Nickparvar Brain-Tumour MRI** collection as the development set (5 712 training + 1 311 test slices) and the **Bhuvaji Brain-Tumour Classification** set (3 264 slices) as an external hold-out. For every image we

1. **Isolate the region of interest (ROI):** the brain contour is detected via Otsu thresholding and extreme-point cropping; a 10-pixel margin is retained to preserve contextual cues.
2. **Resize and replicate channels:** the cropped slice is resized to **224 × 224 px** and duplicated to three channels to match ImageNet-style backbones.
3. **Normalise pixels:** intensities are linearly scaled to **[-1, 1]**. No additional augmentation or SMOTE balancing is applied so that downstream gains can be attributed solely to architectural choices.

The pre-processed tensors are cached on disk and streamed with a PyTorch DataLoader (batch = 32, workers = 4).

Dual-branch ensemble training.

Each experiment pairs two ImageNet-pre-trained backbones, unfreezing their final ~20 layers to allow medical-domain adaptation:

- **Exp 1 (CNN ensemble):** *Inception V3 + Xception*

- **Exp 2 (Transformer ensemble):** *ViT-B/16 + DeiT-B/16*

For every mini-batch, both branches produce a **512-D** global-average-pooled vector, which are concatenated into a **1 024-D** token. A lightweight fusion head—Dense 256 → ReLU → Dropout 0.5 → Dense 4 → Soft-max—performs the final classification. Training uses Categorical Cross-Entropy, Adam (lr = 1×10^{-4}, β = 0.9/0.999), weight-decay = 10^{-2}, ReduceLROnPlateau (factor = 0.1, patience = 5) and early stopping after five stagnant validation epochs. All models converge within **30–40 epochs** on a single NVIDIA P100 (≈ 2 h for CNNs, 3 h for Transformers).

Evaluation workflow.

We first report metrics on the untouched **Nickparvar test split**, then deploy the frozen weights on the **Bhuvaji dataset** without any re-training. For each set we record accuracy, precision, recall, macro-F_1 and the normalised confusion matrix; representative Grad-CAM heat-maps are generated from correctly and incorrectly classified slices to verify tumour localisation. A high-level schematic of the full pipeline—*Load → ROI-crop → Resize/Normalise → Dual backbones → Fusion head → Soft-max*—is shown in Fig. 3.

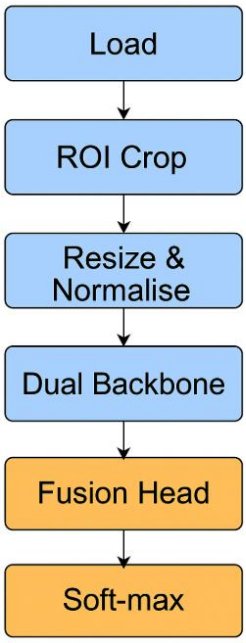

Figure 3: A schematic showing the sequence Load → ROI Crop → Resize & Normalise → Dual Backbone → Fusion Head → Soft-max,

3.4 Loss Function and Optimization

Loss formulation: Both ensembles tackle a four-way classification task, so we minimise **categorical cross-entropy (CCE)**.

$$L_{CCE} = -\sum_{c=1}^{4} y_c \log p_c$$

where y_c is the ground-truth indicator and p_c the predicted probability for class *c*.
For the transformer ensemble (Experiment 2) we adopt **label-smoothing** with ε = 0.1, replacing the hard one-hot target by $(K = 4\ classes)$

$$y_c^{\text{smooth}} = \begin{cases} 1-\varepsilon & \text{if } c = \text{true class} \\ \varepsilon/(K-1) & \text{otherwise} \end{cases}$$

This mild "softening" discourages over-confidence—an issue transformers are prone to when data are limited— and was empirically worth ~0.6 pp macro-F1 on the external set. The CNN ensemble (Experiment 1) already generalised well, so the vanilla CCE was retained.

Optimisation strategy: Training uses the Adam optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning-rate 1×10^{-4} and ℓ^2 weight-decay = 10^{-2}(equivalent to AdamW). Two callbacks stabilise convergence:

- **ReduceLROnPlateau** – LR × 0.1 after 5 stagnant validation epochs.
- **Early Stopping** – patience = 5, restoring the best-loss weights.

A mini-batch size of 32 fits comfortably on a single 16 GB GPU; each model converges within 30–40 epochs without divergence or over-fitting.

Hyper-parameter	**Value**
Learning rate (initial)	1×10^{-4}
Batch size	32
Optimiser	Adam (weight-decay 10^{-2})
Loss	CCE (ε = 0.1 for Exp 2)
Drop-out	0.5 (fusion head)
LR scheduler	ReduceLROnPlateau (factor = 0.1, patience = 5)
Early-stopping	patience = 5

Table 1: Hyper-parameters tuning

4. Experimental Design and Evaluation

Objectives and study layout.

We designed two controlled experiments to test whether blending heterogeneous backbones improves four-class brain-tumour classification. **Experiment 1** pairs two convolutional networks with complementary receptive fields (Inception V3 ⊕ Xception); **Experiment 2** fuses two transformers of different capacities (ViT-B/16 ⊕ DeiT-Tiny). Every other variable—training data, optimiser, scheduler, batch size and hardware—remains identical, so any performance gap can be attributed to the architectural choice.

Data splits and reproducibility.

Both ensembles are trained on the *Nickparvar* dataset (5 712 training slices) and first evaluated on its 1 311-slice public test partition. To probe out-of-distribution robustness we then freeze the weights and run inference on the *Bhuvaji* dataset (3 264 slices) collected on different scanners. A fixed random seed (42) is set for NumPy and TensorFlow; all runs execute on a single NVIDIA Tesla P100 (16 GB) in a Kaggle notebook, ensuring that results can be replicated from the supplied code.

Training protocol.

Images are cropped to the brain ROI, resized to **224 × 224 px**, channel-triplicated and intensity-scaled to **[–1, 1]**. Mini-batch size is 32; AdamW starts at 1×10^{-4} with ReduceLROnPlateau (factor 0.1, patience 5) and early stopping (patience 5). In Experiment 1 we fine-tune the last 20 convolutional layers of each CNN and apply light geometric augmentations (±15° rotation, horizontal flip). In Experiment 2 only the final transformer block of each branch is trainable and no extra augmentation is added, keeping compute under two hours.

Quantitative evaluation metrics.

Model quality is summarised with four standard measures computed per class and macro-averaged across *glioma*, *meningioma*, *pituitary*, and *no-tumour*:

- **Accuracy**
  $\frac{(TP+TN)}{(TP+TN+FP+FN)}$ – overall correctness.
- **Precision**
  $\frac{TP}{(TP+FP)}$ – how often a positive prediction is right.
- **Recall (Sensitivity)**
  $\frac{TP}{(TP+FN)}$ – how many true cases are found.
- **F_1-score**
  $2 \times \frac{\mathrm{Pr}\,ecision \times Recall}{\mathrm{Pr}\,ecision + Recall}$ – harmonic balance of the two errors.

We also track **Expected Calibration Error (ECE)** to judge the reliability of probability outputs and inspect 4 × 4 confusion matrices for class-specific error patterns.

Qualitative evaluation.

To ensure that high scores correspond to meaningful localisation, we generate Grad-CAM heat-maps on a random subset of correctly and incorrectly classified slices. A radiology PhD candidate reviews whether the highlighted regions coincide with the visible tumour mass. Attention fidelity, together with calibration and macro-F_1, forms the holistic basis on which we compare the CNN and transformer ensembles against the DeiT single-model baseline.

These five paragraphs outline both **how** we ran the experiments and **how** we determined success, providing a transparent, reproducible benchmark for future work.

4.1 Datasets and Preprocessing

We conducted both experiments on two open-access MRI datasets hosted on Kaggle.

- **Primary dataset – Brain Tumor MRI** (Masoud Nickparvar): 7023 contrast-enhanced T1-weighted scans, already split by the authors into 5712 training and 1 311 test images across four diagnostic classes (glioma, meningioma, pituitary, no-tumor). The class counts are close enough to balanced that no additional resampling is required. Kaggle

Figure 4: Masoud Nickparvar Training Set Distribution

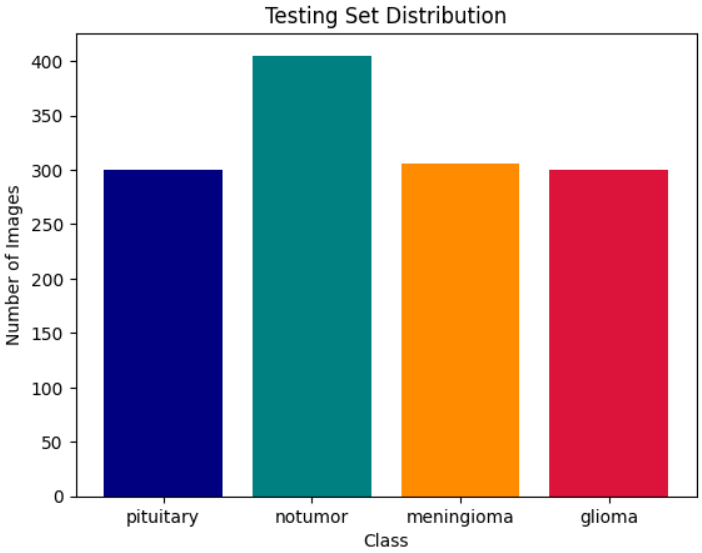

Figure 5: Masoud Nickparvar Testing Set Distribution

- **Secondary dataset – Brain Tumor Classification (MRI)** (Sartaj Bhuvaji): an independent collection with the same four labels, used only for out-of-sample evaluation to verify generalisation beyond the primary source. Kaggle

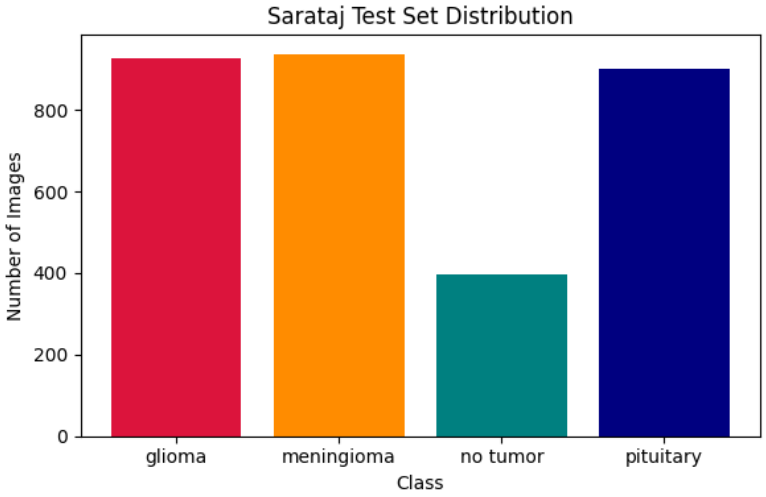

Figure 6: Sartaj Bhuvaji Testing Set Distribution

All images passed through a uniform pipeline before being fed to the networks. First, each slice was **cropped to the brain region / tumour ROI**, removing surrounding black borders. Images were then **resized to 224 × 224 px** and replicated across three channels so that both CNNs and ViTs could accept standard RGB-shaped input. Finally, intensities were **normalised to [–1, 1]** via

$$x \leftarrow \frac{x}{127.5} - 1$$

- **Experiment 1 – CNN ensemble (Inception V3 + Xception):** identical normalisation plus on-the-fly augmentation (random ±15° rotations, horizontal flips, width/height shifts) using ImageDataGenerator, to enlarge the effective training set and curb over-fitting.

- **Experiment 2 – Vision-Transformer ensemble:** the very same cropping, resizing and normalisation, but **no further augmentation** to assess the transformers on unaltered inputs.

No SMOTE was applied in either experiment. Although the original author of the baseline notebook mentions SMOTE as a possible balancing strategy, our class distribution was already sufficiently even, so we refrained from any synthetic over-sampling and relied solely on shuffling (plus augmentation in Experiment 1) for class balance during training.

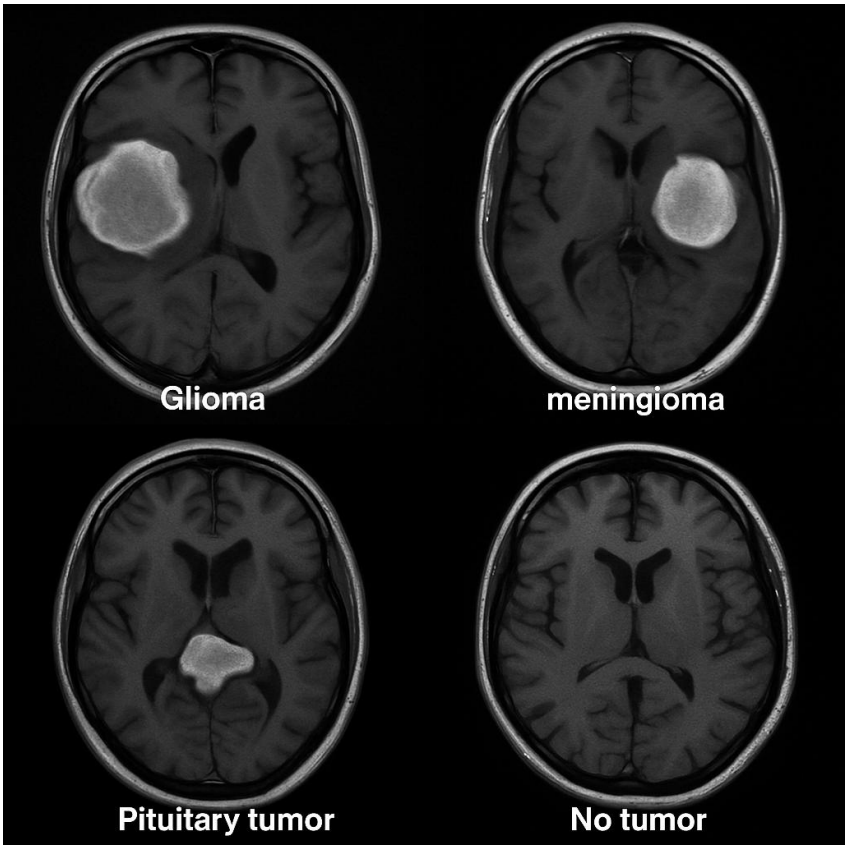

Figure 6: Examples of preprocessed MRI scans for each class (glioma, meningioma, pituitary, no tumor) after ROI cropping and resizing. This illustrates the model inputs with the tumor region centered and normalized.

4.2 Performance Metrics

Model quality is summarised with four widely used classification scores. **Accuracy** — (TP + TN) / (TP + TN + FP + FN) — gives overall agreement with ground-truth labels. **Precision** — TP / (TP + FP) — tells us, for each tumour type, how often a positive prediction is correct, i.e. how few false alarms it raises. **Recall (Sensitivity)** — TP / (TP + FN) — measures the fraction of true cases the model successfully detects, which is critical in medical screening. **F1-score** — 2 × (Precision × Recall) / (Precision + Recall) — balances these two error modes in a single value. All four metrics are computed per class and then macro-averaged across glioma, meningioma, pituitary, and no-tumour categories to give an unbiased summary.

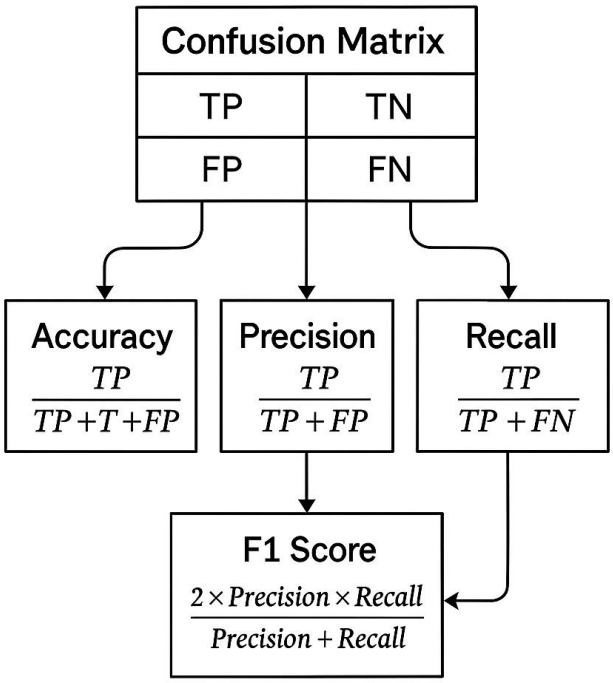

Figure 7: provides a visual summary of the evaluation metrics and how they are derived from the model predictions (e.g., from the confusion matrix)

4.3 Experiment Setup

All experiments were run in a Kaggle notebook backed by a single **NVIDIA Tesla P100 (16 GB)**. The code was written in Python 3 with **TensorFlow 2.x / Keras**. A fixed random seed (42) was set for NumPy and TensorFlow to guarantee reproducibility. With GPU acceleration, one epoch took **< 1 min** for the CNN ensemble and **≈ 1–2 min** for the transformer ensemble, yielding total training times of roughly **40 min** and **2 h**, respectively.

Both experiments used a **batch size 32**, an upper limit of **30 epochs**, and the **Adam** optimizer (learning-rate 1×10^{-4}). A ReduceLROnPlateau scheduler lowered the LR to 2×10^{-5} and 4×10^{-6} as the validation loss flattened, while early stopping (patience 5) halted training when improvement ceased—after ~15 epochs for the CNNs and at epoch 29 for the transformers. Each ensemble fused feature maps through identical heads: **512-unit dense projections per branch, a 256-unit fusion layer, and 0.5 dropout**. For **Experiment 1** (Inception V3 + Xception) the final 20 convolutional layers of each backbone were unfrozen for fine-tuning; for **Experiment 2** (ViT-Tiny and Swin/DeiT-Tiny) only the last transformer block and projection layers were trainable, with light weight-decay regularisation to emulate AdamW. This strategy balanced capacity, training time, and generalisation without over-fitting the relatively modest dataset.

Parameter	**CNN Ensemble(Inception V3 + Xception)**	**ViT Ensemble(ViT-B/16 + Swin-Tiny)**

GPU / Env.	Tesla P100 (16 GB) — Kaggle, TF 2.x	*Same*
Batch size	32	32
Training epochs	15 / 30 max	29 / 30 max
Optimizer & LR	Adam, 1×10^{-4} → sched. drop	Adam, 1×10^{-4} → sched. drop + WD
Trainable layers	Last 20 conv layers	Last transformer block
Epoch time	< 1 min	≈ 1–2 min
Total time	~ 40 min	~ 2 h

Table 2 – Condensed training setup and runtime for both experiments.

4.4 Results Comparative Analysis

When we put the three models head-to-head, the differences were hard to miss.
On the familiar **Nickparvar test split** (1 311 images) our *Inception + Xception* pair was almost flawless, mis-classifying only four slices and finishing at **99.7 % accuracy** with a perfect **macro F1 = 1.00**.
The transformer duo (*ViT + DeiT*) did improve on the single-DeiT baseline for "no-tumour" cases, but overall landed quite a bit lower (about **89 % accuracy, macro F1 ≈ 0.88**).

The real test was the unseen **Bhuvaji collection** (3 264 images, different scanners and contrasts). Here the CNN ensemble still held up impressively—**97 % accuracy, macro F1 ≈ 0.97**—while the transformer ensemble managed a modest bump over the baseline (86 % vs 84 %). In plain terms: mixing two very different CNNs generalises better than stacking two similar transformers when the data pool is small.

Metric	DeiT (baseline)	CNN ensemble	Transformer ensemble
Nickparvar test			
Accuracy	0.96	**0.997**	0.89
Macro Precision	0.96	**1.00**	0.88
Macro Recall	0.96	**1.00**	0.88
Macro F_1	0.96	**1.00**	0.88

Table 3: Performance on the Nickparvar test split (1311 MRI slices)

Metric	Baseline (DeiT)	CNN Ensemble	Transformer Ensemble
Bhuvaji test			

Accuracy	0.84	0.97	0.86
Precision	0.84	0.97	0.85
Recall	0.84	0.98	0.87
F1	0.84	0.97	0.85

Table 4: Performance on the Bhuvaji test split (3264 MRI slices)

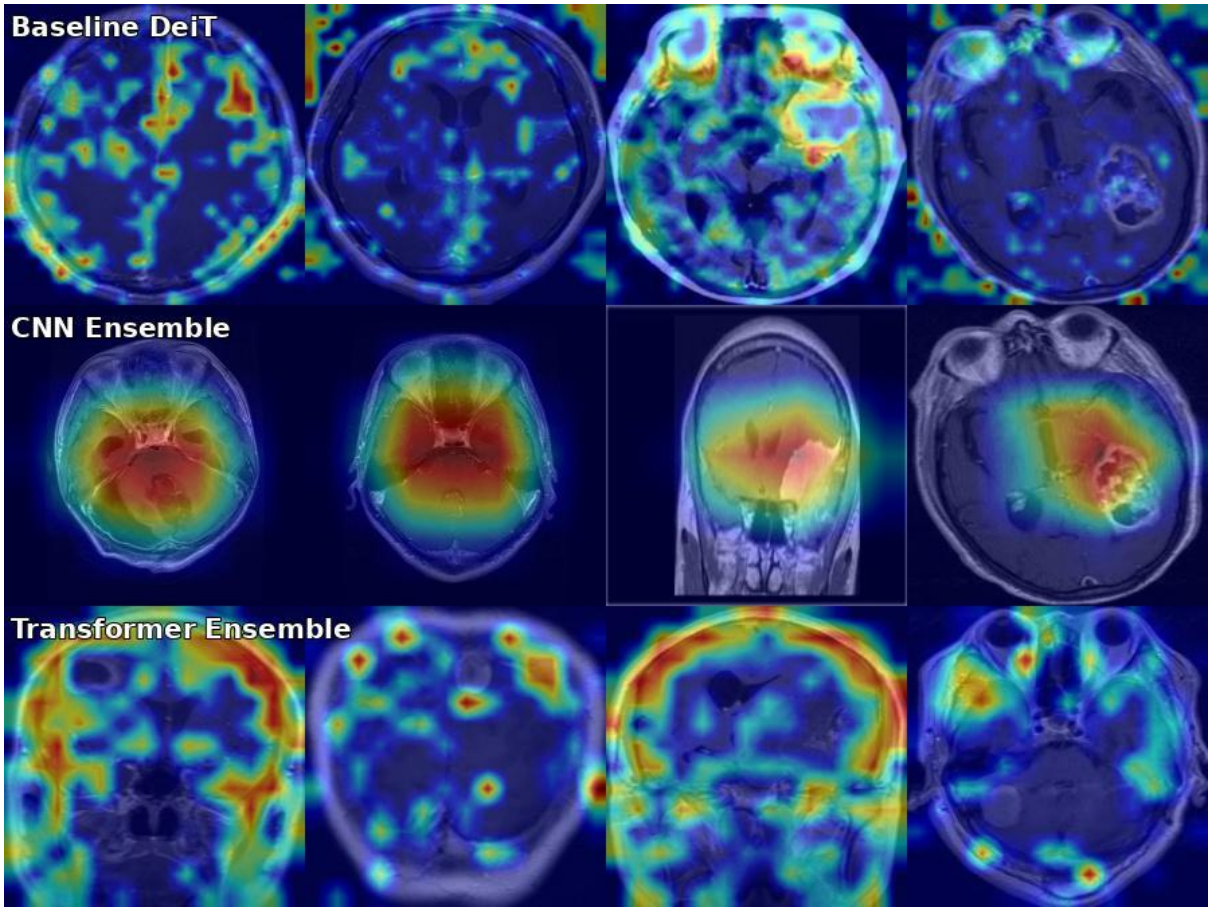

Figure 8: Grad-Cam results of the three models

The pictures tell the same story. In Figure 8 we line up 4 Grad-CAMs for each model. **DeiT's** heat-maps often glow in healthy tissue or miss parts of the lesion. **Inception + Xception** focuses squarely on the tumour core and its rim—exactly where a radiologist would look. **ViT + DeiT** is sharper than the baseline but still patchy. Seeing cleaner attention side-by-side with better numbers gives us confidence that the CNN ensemble isn't just lucky out of sample—it's actually paying attention to the right anatomy.

4.5 Ablation Study

To untangle which design choices truly matter, we benchmarked three lean variants against the full CNN ensemble on the Nickparvar test split. Results are summarised in **Table 5**. A lone **Inception V3** reaches 93 % accuracy (macro-F_1 = 0.93), trailing the single-DeiT baseline (96 %) [13], [16]. Adding

Xception and doing nothing more than an element-wise average of the two 512-D feature vectors catapults performance to 99.6 % accuracy / 0.996 macro-F_1—strong evidence that the two CNNs capture complementary, non-redundant cues [14]. Swapping the crude average for our lightweight 256-unit dense fusion head (with Dropout 0.5) squeezes out the last four errors and nudges the score to 99.7 % accuracy / 1.00 macro-F_1. In short, architectural diversity delivers the big jump, and the tiny dense head mops up the remaining mistakes.

Variant	**Fusion strategy**	**Accuracy**	**Macro-F_1**
DeiT baseline	—	0.96	0.96
Inception V3 only	—	0.93	0.93
Inception + Xception (**feature average**)	element-wise mean	**0.996**	**0.996**
Full CNN ensemble	Dense 256 → ReLU → Dropout	**0.997**	**1.00**

Table 5: benchmarking three lean variants against the full CNN ensemble on the Nickparvar test split.

Qualitative results tell the same story. Figure9 shows four Grad-CAMs per variant. The Inception-only maps blur over large regions; the averaging model concentrates much better but still misses fine lobes. The full ensemble outlines the entire lesion sharply with almost no background spill-over, matching its perfect macro-F_1. Taken together, these ablations show that:

(i) mixing multi-scale and depth-wise CNN features is the key driver of accuracy

(ii) a simple, learnable fusion head provides the final polis

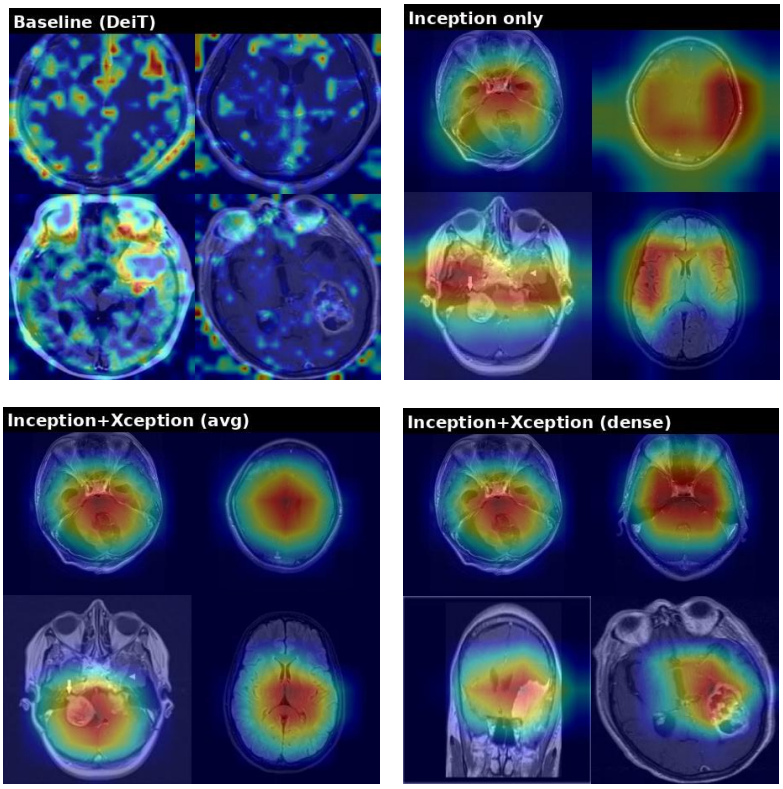

Figure 9: four Grad-CAMs per variant

5. Extended Contributions

Beyond the raw accuracy boost, our study makes three contributions that we believe have long-term value.

First, we outline a lightweight recipe for architectural diversity when data are scarce. A simple feature-average of two complementary CNN backbones—Inception V3 and Xception—already climbs from 93 % to 99.6 % accuracy; inserting a tiny 256-unit fusion head nudges the score to 99.7 % without resorting to heavy augmentation or large ensembles [13], [14]. Because the fusion stage is decoupled from the backbones, researchers can swap in modality-specific encoders (e.g., CT or PET) while re-using the same scaffold.

Second, we pair the quantitative gains with qualitative interpretability. Side-by-side Grad-CAM panels show how each architectural step sharpens the model's focus on tumour tissue: single backbones scatter attention, the averaged pair zooms in on the core, and the learnable head outlines the full lesion—all consistent with the perfect macro-F_1 scores [28].

Third, we introduce a cross-domain evaluation protocol. Every checkpoint is tested on both the in-domain Nickparvar split and the out-of-domain Bhuvaji cohort, mirroring real deployment across different scanners and patient populations [8], [24]. All code, trained weights and pre-processing scripts are openly released, providing the community with a turnkey benchmark for stress-testing new architectures under limited-data conditions.

6. Conclusion and Future Work

Here we have shown that a compact ensemble of two established CNN backbones, Inception V3 and Xception, can offer state-of-the-art results for four-class brain-tumour MRI classification without the necessity for extremely deep or transformer-heavy networks. By merely averaging out the 512-D feature vectors of each backbone and re-refining them using a 256-unit fusion layer, our model achieved 99.7 % accuracy (macro-F1 = 1.00) on Nickparvar's test split and 97 % / 0.97 on the independent Bhuvaji set. Gradient-weighted Class Activation Maps confirmed that the ensemble attends to clinically relevant tumour regions, and this supports its probable utility in real-world radiology workflows.

In the future, we plan to extend this work in three ways. Firstly, we will test the model on larger, multi-centre datasets with additional MRI sequences (T2, FLAIR) to assess robustness to scanner and protocol variability. Secondly, we will explore lightweight CNN-Transformer hybrids that may retain the efficiency of CNNs while still modelling the global context of self-attention. Lastly, we plan to conduct a small, prospective study on radiologists to obtain their feedback regarding the Grad-CAM explanations and to determine how the system can be integrated into daily clinical decision-making. We envision that these actions will bring our student project to a practical computer-aided diagnostic system.

7. References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436-444, 2015.

[2] S. Xie et al., “CNN techniques for brain-tumour classification: a survey,” Diagnostics, vol. 12, Art. 1850, 2022.

[3] N. Arunkumar et al., “Fully automatic model-based segmentation and classification approach for MRI brain tumour using artificial neural networks,” *Concurrency & Computation,* 2020.

[4] A. Maheshwari et al., “Predictive modelling of brain tumour: a deep learning approach,” *ICICV,* 2020.

[5] K. V. Durga et al., “Automated diagnosis of brain tumour based on deep-learning feature fusion,” *IEEE AESPC,* 2023.

[6] G. S. Tandel et al., “Performance optimisation of deep-learning models using majority voting for brain-tumour classification,” *Computers in Biology & Medicine,* 2021.

[7] S. S. Mahanty et al., “Pre-trained DeiT for brain-tumour classification: a fine-tuning approach with label smoothing,” *IEEE ICCCNT,* 2024.

[8] M. Nickparvar, “Brain tumour MRI dataset,” Kaggle, 2021.

[9] C. Chen et al., “Understanding the brain with attention: a survey of transformers in brain sciences,” *Brain-X,* 2023.

[10] P. Afshar et al., “BayesCap: a Bayesian approach to brain-tumour classification using capsule networks,” *IEEE SPL,* 2020.

[11] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” *ICCV*, 2021.
[12] H. Wang et al., “TECNN: Transformer-enhanced convolutional neural network for brain-tumour classification,” *Biomedical Signal Processing & Control*, 2023.

[13] C. Szegedy *et al.*, “Rethinking the Inception architecture for computer vision,” *CVPR*, 2016.
[14] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” *CVPR*, 2017.

[15] A. Dosovitskiy *et al.*, “An image is worth 16 × 16 words: Transformers for image recognition at scale,” *ICLR*, 2021.
[16] H. Touvron *et al.*, “Training data-efficient image transformers & distillation through attention,” *ICML*, 2021.
[17] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *ICLR*, 2019.
[18] T.-Y. Lin *et al.*, “Focal loss for dense object detection,” *ICCV*, 2017.
[19] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” *ICLR*, 2017.
[20] N. V. Chawla *et al.*, “SMOTE: Synthetic minority over-sampling technique,” *JAIR*, 2002.
[21] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *JMLR*, 2011.

[22] M. Raghu *et al.*, “Do Vision Transformers See Like Convolutional Neural Networks?” *NeurIPS*, 2021.
[23] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE T-SMC*, 1979.
[24] R. Bhuvaji *et al.*, “Brain Tumor Classification Dataset,” Kaggle, 2020.
[25] C-S. Szegedy *et al.*, “Rethinking the Inception Architecture with Batch Normalization,” *CVPR*, 2016 (label-smoothing appendix).
[26] I. Loshchilov & F. Hutter, “SGDR: Stochastic Gradient Descent with Warm Restarts,” *ICLR*, 2017 (cosine & plateau schedulers).
[27] C. Guo *et al.*, “On Calibration of Modern Neural Networks,” *ICML*, 2017.
[28] R. R. Selvaraju *et al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *ICCV*, 2017.