International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I7P119

Volume 74 Issue 7, 281-309, July 2026

Original Article

Reliability and Explainability Analysis of CNN and Vision Transformer Models for Brain Tumor MRI Classification

Anuj Gupta[1], Anita[2], Manish Gupta[3]

[1,2]*Department of CSE & IT, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India.*
[3]*Department of Radiation Oncology, Indira Gandhi Medical College, Shimla, Himachal Pradesh, India.*

[1]*Corresponding Author : nsanujgupta@gmail.com*

Received: 14 February 2026 Revised: 10 June 2026 Accepted: 18 June 2026 Published: 28 July 2026

Abstract - *The clinically feasible explanations and the high performance in classifications are not only strong screening outcomes of the Magnetic Resonance Imaging (MRI) on brain tumors, but also trustworthy confidence estimates and explanations. The present paper gives a single-run benchmark of three convolutional neural networks (VGG16, ResNet50, EfficientNetB0) and two families of vision transformers (ViT-Base and Swin-Base) in a 100-epoch training scheme in four-class brain MRI classification (glioma, meningioma, pituitary, and no-tumor). It has accuracy, macro-precision/recall/F1, micro-average ROC: AUC, per-class, and confusion matrix reports. In order to measure reliability, it determines calibration by on-assessment of reliability diagrams and Estimated Calibration Error (ECE). To evaluate the interpretability, Grad-CAM images are given to all five models using post-processing (thresholding and smoothing) to identify class-discriminative areas. In addition to discrimination, the assessment presents the differences in confidence, reliability and the quality of explanation among architects as noteworthy, indicating that the accuracy level alone is insufficient to use a system in clinical practice. The outcomes of this work bring forward the need to be reliability-conscious and responsive towards explainability validation before the process is transformed into real neuroimaging practices. The modern backbones (EfficientNetB0 and Swin) are more discriminative and better calibrated models than VGG16, and saliency explainability is also more correlated with pathology, as opposed to classical baselines. It puts the findings in context of the more current state-of-the-art, in terms of summarizing those recently advanced variants of transformer that can be tuned to better calibration output (up to 98.9% accuracy and ECE as low as 0.023) when using multi-scale patch embeddings, feature calibration modules and selective attention.*

Keywords - *Brain tumor classification, MRI, Convolutional neural networks, Vision transformer, Swin transformer, Calibration, Reliability, Grad-CAM, Explainable AI.*

1. Introduction

Brain tumors represent a serious class of neurological disorders in which early and accurate diagnosis plays an important role in treatment planning, surgical decision-making, radiotherapy design, and patient prognosis. Magnetic Resonance Imaging (MRI) is widely used for an assessment of brain tumors due to its exceptional soft-tissue contrast relative to various other imaging techniques. Nonetheless, the manual interpretation of MRI data is labor-intensive and may differ among radiologists due to variations in expertise, visual perception, and tumor heterogeneity. Consequently, deep learning-based computer-aided diagnosis has emerged as a significant research focus for the automated classification of brain tumors and decision support [1, 2]. “Convolutional Neural Networks (CNNs)” have been frequently used in this domain because they can learn hierarchical spatial information, whereas transformer-based models have lately gained popularity because they can capture long-range contextual dependencies. Nevertheless, high classification accuracy alone is insufficient for clinical adoption because a model may still produce overconfident predictions, poorly calibrated probability scores, or explanations that do not correspond to clinically meaningful tumor regions [3].

In recent years, Vision Transformers (ViTs) and hierarchical transformer architectures that involve Swin Transformer have demonstrated competitive performance in visual recognition and medical imaging tasks [4-6]. Their self-attention mechanism enables the modelling of global relationships that may be useful for capturing tumor morphology, surrounding tissue patterns, and contextual features. However, recent studies and reviews on medical artificial intelligence emphasize that clinical translation requires more than aggregate accuracy [7, 8]. A reliable

diagnostic model should provide not only correct predictions but also trustworthy confidence estimates and interpretable visual evidence. This requirement is particularly important in brain tumor classification, where a confident false prediction may delay diagnosis, misguide treatment planning, or create unnecessary clinical intervention.

From a clinical perspective, all diagnostic errors do not carry the same level of risk. A false-negative prediction may delay further investigation, while an incorrect tumor subtype prediction may affect surgical planning, radiotherapy strategy, or adjuvant treatment selection. Therefore, automated models intended for neuro-oncology should not be evaluated only through predictive accuracy. They must also be assessed in terms of uncertainty, calibration, robustness, and explanation quality. In addition, brain tumor MRI classification is affected by several domain-specific challenges, including variations in tumor shape, intensity, boundary appearance, scanner protocol, slice thickness, and acquisition conditions. These factors can influence the generalization behaviour of deep learning models and must be considered when evaluating their clinical usefulness.

To start with, it is common to find that tumors may have heterogeneous morphological and intensity patterns even when belonging to the same diagnostic category, and peripheral edema may be accompanied by necrosis, so as to have an influence on the appearance of the boundaries. Second, acquisition settings (scanner vendor, field strength, slice thickness and contrast protocol) also result in non-trivial variability, which may also undermine generalization if they are not explicitly considered. Third, some clinically promising errors are asymmetric, such as the confusion of Glioma with meningioma might result in divergent downstream work-up and treatment planning, but a sure false prediction of the form no-tumor may be deferring.

These factors make a single aggregate performance score inadequate for clinical interpretation. A clinically meaningful evaluation should examine:

(i) class-wise performance and failure modes,
(ii) whether predicted confidence values are reliable indicators of correctness, and
(iii) whether visual explanations are aligned with plausible lesion-related evidence.

This broader evaluation is necessary because two models with similar accuracy may differ substantially in calibration, uncertainty behaviour, and explanation quality.

Recent research has increasingly shown that deep neural networks, including CNNs and transformer-based models, may produce overconfident probability estimates when trained using standard optimization objectives. This miscalibration is problematic in clinical decision-support systems because confidence scores may be interpreted as indicators of diagnostic certainty. A poorly calibrated model may assign high confidence to an incorrect prediction, thereby reducing the likelihood of expert review. Reliability analysis through calibration curves, Expected Calibration Error, and uncertainty estimation is, therefore, essential for determining whether model outputs can support triage, prioritization, or second-reader workflows.

1.1. Motivation

A clinically useful brain tumor MRI classifier should satisfy three practical requirements. First, it must perform well on all tumor classes and not just on the majority or easier-to-see classes. Second, its confidence scores ought to be dependable to facilitate triage, referral and expert-review decisions. Third, it should emphasize image areas that are in line with plausible radiological findings in its explanations. The importance of these requirements is that a slightly more accurate model with a poor calibration could be unsuitable in safety-critical applications, whereas a slightly less accurate model with good confidence estimates and clear evidence could be useful.

Available literature on brain tumor MRI classification usually claims high accuracy when CNNs, transfer learning models, or transformer-based models are used. Nevertheless, a lot of these studies do not collectively explore calibration, uncertainty, and quality of explanation. Specifically, a comparison between CNN and transformer models based on a standard evaluation protocol is still scarce.

In addition, the majority of 2D slice-based research lacks adequate coverage of their limitations as compared to multi-sequence MRI, volumetric setting, external validation, domain shift, and human-centered assessment. This sets a disconnect between high-performing experimental models and clinically trustworthy decision-support systems [8, 9].

The motivation of this study is to address this gap by evaluating CNN and transformer models through a broader reliability-and-explainability framework. The paper contrasts VGG16, ResNet50, EfficientNetB0, ViT-Base, and Swin-Base in four-class brain tumor MRI classification with a single training regime.

The assessment is complemented by the common classification measures, such as confusion-matrix analysis, reliability diagrams, Expected Calibration Error, MC Dropout-based uncertainty summaries, and Grad-CAM-based visual explanation.

This design allows a more balanced evaluation of model behaviour as it looks at not just the frequency of successful prediction by a model, but also the confidence with which the model predicts and the clinical plausibility of its visual evidence.

1.2. Contributions

This study provided these key contributions:

(i) A unified comparative evaluation of three CNN backbones and two transformer-based models for four-class brain tumor MRI classification.
(ii) A multi-dimensional assessment framework covering accuracy, macro-averaged precision, recall, F1-score, micro-average ROC-AUC, class-wise performance, and confusion-matrix behaviour.
(iii) Conduct reliability analysis using reliability diagrams and Expected Calibration Error to verify predicted probability have empirical accuracy.
(iv) Uncertainty-oriented analysis using MC Dropout summaries to discuss confidence behaviour and its relevance to deferral-based clinical workflows.
(v) Grad-CAM-based interpretability analysis with thresholding and smoothing to compare whether model attention is visually aligned with plausible tumor-related regions.
(vi) A clinical translation discussion covering robustness to domain shift, multi-sequence MRI, external validation, quantitative explainability validation, and human-centered evaluation.

2. Related Work

Over the past decade, deep learning-based brain tumor MRI categorization has improved. CNNs may extract local spatial information, including edges, textures, shape patterns, and lesion boundaries, therefore early investigations focused on them. Recent CNN-based transfer-learning studies, including EfficientNetB0-based brain tumor multi-classification accuracy in brain tumor recognition tasks [10]. However, many CNN-based studies primarily report accuracy, precision, recall, and F1-score, while limited attention is given to probability calibration, uncertainty estimation, external validation, and explanation reliability. This creates a gap between experimental classification performance and clinical trustworthiness.

A major limitation of conventional CNN-based evaluation is that the predicted softmax score is often treated as a direct measure of confidence. In practice, this assumption may not hold because deep neural networks can be highly accurate and still poorly calibrated. An overconfident, incorrect prediction is particularly risky in clinical decision support because it may reduce the likelihood of expert review. Thus, reliability diagrams and Expected Calibration Error are frequently used in medical AI [8, 9]. Despite this need, only a limited number of brain tumor MRI classification studies systematically compare calibration behaviour across different CNN and transformer architectures.

Vision Transformers use self-attention to model long-range dependencies across image regions as an alternative to CNNs. Hierarchical variations like Swin Transformer use shifted-window attention to increase computing efficiency and multi-scale representation learning, while ViT models divide an image into patches and learn global relationships [4, 5]. In medical imaging, transformer-based models that involve MedViT, TransUNet, and UNETR have shown the potential of attention-based representation learning for classification and segmentation tasks [6, 12, 13]. In brain tumor analysis, recent transformer-based studies have reported strong performance, particularly when multi-scale attention, hierarchical feature extraction, and domain-specific fine-tuning are used. However, most studies still focus on accuracy improvement and do not sufficiently analyze whether transformer models provide better calibration, uncertainty behaviour, or explanation quality than CNNs.

Another major medical imaging research area is explainable AI. Visualizing picture regions that affect model predictions with Grad-CAM and saliency-based approaches is popular [11, 21]. Visual explanations can help brain tumor MRI classification models focus on lesion-relevant regions or irrelevant background patterns. Nevertheless, saliency maps are often presented only as qualitative examples. Their interpretation may vary depending on layer selection, preprocessing, post-processing, and visualization settings. Recent explainability studies therefore recommend cautious interpretation of saliency maps and encourage quantitative validation using expert annotations, overlap scores, Dice coefficient, IoU, or localization agreement measures [7, 11, 22].

Another important limitation in existing brain tumor MRI classification research is the frequent use of 2D single-sequence slices. Although 2D slice-based models are computationally efficient and useful for benchmarking, they may not fully capture volumetric tumor extent, inter-slice continuity, edema patterns, necrosis, enhancement behaviour, or multi-sequence radiological characteristics. Current clinical neuro-oncology assessment often relies on multiple MRI sequences and, where available, clinical information. Therefore, multi-modal and multi-sequence approaches using T1, T1-contrast, T2, FLAIR, or clinical metadata can provide richer diagnostic context than single-sequence 2D models. Recent studies using 3D and multi-modal MRI have demonstrated the value of volumetric and sequence-level information, but such approaches are not always directly comparable with lightweight 2D classifiers due to differences in data availability, computational cost, and annotation requirements [17, 19].

Uncertainty-aware deep learning has also gained importance for safe medical AI deployment. Methods that involve Bayesian neural networks, Monte Carlo Dropout, deep ensembles, evidential learning, and post-hoc calibration can estimate predictive uncertainty and help identify cases that should be referred for expert review [20, 23]. These techniques are particularly important for ambiguous MRI slices, low-quality images, out-of-distribution inputs, and

borderline tumor appearances. However, uncertainty-aware methods are still not routinely benchmarked in brain tumor MRI classification studies, especially in direct comparisons between CNNs and transformer-based models.

The above literature indicates that the field has achieved strong progress in classification accuracy, but important gaps remain in reliability, uncertainty, explainability validation, and clinical deployment readiness. Existing studies often evaluate either CNNs or transformers in isolation, rely heavily on aggregate accuracy, and provide limited discussion of calibration, domain shift, multi-center robustness, or human-in-the-loop evaluation. This study addresses this gap by comparing sample CNN and transformer models under a shared protocol and analyzing discrimination, calibration, uncertainty, and Grad-CAM-based explanation behavior. This reliability-oriented benchmark complements architecture-focused studies by emphasizing whether model outputs are sufficiently trustworthy and interpretable for future clinical decision-support workflows.

2.1. Recent Advances beyond Baseline ViTs

Recent works have further developed baseline Vision Transformers to enhance brain tumor MRI classification with multi-scale representation learning, feature calibration, hierarchical attention, and efficient token selection. Such developments apply to the current study as they demonstrate that the performance of transformers is not only determined by the application of self-attention, but also by the manner in which local, global and multi-scale information is incorporated. The utilization of hierarchical attention and multi-scale patch embeddings is one of the directions. Multi-scale transformer designs seek to simultaneously extract image patches at various resolutions and fuse the image patches with attention-based fusion to capture tumors of varying sizes. The designs are applicable to brain tumor MRI since the appearance of tumors can differ significantly between classes and even between cases. As an illustration, hierarchical and multi-scale attention processes have been found to enhance classification accuracy and calibration through the ability of the model to focus on fine-grained lesion texture, as well as on a large anatomical context [15].

The other current trend is feature calibration and selective cross-attention. The feature calibration modules are aimed at perfecting the token representations prior to final classification, and selective cross-attention may integrate local and global information in a more efficient way. Such methods are especially applicable to medical imaging since the model decision can otherwise be biased by irrelevant background structures, scanner-dependent patterns or non-lesion areas. Recent calibration-sensitive transformer models have shown good classification results on brain tumor datasets, indicating that the refinement of features based on calibration can enhance both classification and reliability of the decisions [16].

Brain tumor MRI classification using fine-tuned transformer models has also been explored. These methods do not need to train transformers initially, but rather exploit large-scale trained transformer backbones and reconfigure them to medical imaging tasks via transfer learning. Transformer representations have demonstrated competitive results with conventional CNN baselines, showing that fine-tuned ViT models can be useful even in the case of relatively limited medical datasets [17]. Comparative research of transformer families shows that hierarchical models like Swin Transformer balance accuracy, computational efficiency, and inference speed [18].

In addition to single-sequence 2D MRI classification, multi-modal and multi-sequence MRI inputs have also been studied recently. Complementary MRI sequences that are commonly used in clinical brain tumor assessment include T1, contrast-enhanced T1, T2 and FLAIR since each sequence indicates different tissue attributes. Models that use multi-sequence or volumetric data can thus give more information than single-slice models. To illustrate, transformer-based models with multi-modal MRI data have also demonstrated high classification accuracy on the basis of complementary information in various imaging sequences [19]. Nevertheless, these techniques frequently involve more complicated preprocessing, increased computing resources, and more comprehensive datasets than 2D slice-based benchmarks.

Another key finding that can be used to achieve clinically safe deployment is uncertainty-aware learning. “Monte Carlo Dropout”, “deep ensembles, “Bayesian neural networks”, and “post-hoc calibration” techniques can provide extra information regarding the confidence of the prediction and the uncertainty of the model [22]. They can be used to detect ambiguous or out-of-distribution cases, which must be sent to radiologists instead of being accepted automatically. Uncertainty-aware brain tumor MRI classification can improve selective prediction, confidence-based triage, and human-in-the-loop safety.

In general, the recent versions of transformers demonstrate that the high classification accuracy could be enhanced with the help of multi-scale attention, feature calibration, selective attention, and multi-modal learning. But the literature suggests too that the improvement of accuracy does not necessarily result in clinical reliability. Numerous high-performing studies continue to perform scarce analysis on the calibration, uncertainty, explainability validation, domain shift and user-centered evaluation. Thus, the current research adopts baseline CNN and transformer models as exemplary backbones and assesses them within a more comprehensive reliability-and-explainability framework. This helps establish a clinically relevant benchmark before introducing more complex transformer variants or multi-modal extensions.

3. Materials and Methods

3.1. Dataset and Preprocessing

This research employs a four-class brain MRI dataset of T1-weighted axial MRI slices categorized as glioma, meningioma, pituitary tumor, and no tumor. The collection comprises 7,023 pictures sourced from publically accessible brain tumor MRI archives. The class distribution is very equitable, with glioma, meningioma, pituitary tumor, and no-tumor images comprising around 23.1%, 23.4%, 25.0%, and 28.5% of the sample, respectively [15].

All images had been scaled to 224 × 224 pixels to match CNN and transformer backbone input requirements. Pre-training pixel intensities were standardized. A stratified split divided the dataset into training, validation, and test subsets to maintain class distribution. The identical division was employed across all models to guarantee an equitable comparison.

Light data augmentation was used during training to enhance generalization without harming anatomical plausibility. The augmentation functions were the horizontal flipping, small-angle rotation, random zooming, and contrast adjustment. No augmentation of validation and test images. Unnecessarily violent changes were not employed since they can alter tumor morphology or change image patterns of clinical significance.

Though the dataset itself is handy when it comes to benchmarking controlled models, it should be noted that it is founded on 2D single-sequence MRI slices. Thus, the dataset is not a complete representation of volumetric tumor continuity, multi-sequence MRI features, scanner-vendor differences, and multi-center acquisition differences. This is a limitation that is specifically addressed in the robustness and future clinical translation.

3.2. Models Evaluated

Five representative deep learning architectures were evaluated: VGG16, ResNet50, EfficientNetB0, ViT-Base, and Swin-Base. VGG16, ResNet50, and EfficientNetB0 were selected as CNN-based models because they represent classical, residual, and compound-scaled convolutional design strategies. ViT-Base and Swin-Base were selected as transformer-based models because they represent global self-attention and hierarchical shifted-window attention mechanisms, respectively.

Each model was initialized with ImageNet pre-trained weights and customized for four-class brain tumor MRI classification using a task-specific classification head. CNN classifiers used global average pooling, fully connected layers, and softmax output layers. For transformer models, the final embedding representation was connected to a classification layer with four output neurons. The use of pre-trained weights was intended to improve convergence and reduce the risk of overfitting on a moderate-sized medical imaging dataset.

3.3. Training Protocol

All models were trained under a common experimental protocol to ensure comparability across architectures. The AdamW optimizer was used to train the models for up to 100 epochs, with an initial learning rate of 1×10^{-4} and weight decay of 0.05. Using a cosine learning-rate plan with warm-up stabilized training. Reduced classification overconfidence was achieved with label smoothing at 0.1. CNN and transformer models varied batch sizes for memory reasons.

The preprocessing, dataset split, loss function, optimizer family, and evaluation metrics remained unchanged across all of the models, though. The best model checkpoint was chosen in terms of the validation performance and tested on the independent test subset.

Experiments were repeated with many random seeds wherever computationally possible to minimize the impact of random initialization and variability in data ordering. Mean values and standard deviations are recommended for reporting model performance in the final tabular results. This helps determine whether observed differences between architectures are stable rather than caused by a single favourable run.

3.4. Evaluation Metrics

"Accuracy, "" precision, "" recall, "" F1-score, and "micro-ROC-AUC" assessed model performance. The macro-averaged metrics were added since they give an equal picture of the performance of all the classes and are not dominated by the majority class. To analyze the patterns of errors by class, especially those between glioma and meningioma, confusion matrices were created with potential clinical implications.

For each class, precision, recall, and F1-score were also calculated to identify whether specific tumor categories were more difficult to classify. This class-wise evaluation is important because a model with high overall accuracy may still perform poorly for one clinically important class.

3.5. Reliability and Calibration

Reliability was evaluated using calibration curves and Expected Calibration Error. Calibration analysis examines whether predicted confidence values correspond to empirical correctness. In clinical decision-support systems, this is important because a confidence score may influence whether a case is accepted, flagged, or referred for expert review.

Let $\hat{p}_i$ denote the confidence score of the predicted class for the sample i, and let $\hat{y}_i$ and y_i denote the predicted and true labels, respectively. The predictions are divided into M confidence bins B_m. The accuracy and confidence of each bin are computed as:

$$acc(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} 1\,(\hat{y}_i = y_i) \quad (1)$$

$$conf(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{p}_i \quad (2)$$

The Expected Calibration Error is then calculated as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|, \quad (3)$$

In the above equations, $\hat{p}_i$represents the predicted confidence score of the model for the i^{th}test sample. This confidence score is usually the maximum softmax probability assigned to the predicted class. The term $\hat{y}_i$denotes the predicted class label, whereas $y_i$represents the actual ground-truth class label. The indicator function $\mathbf{1}(\hat{y}_i = y_i)$returns a value of 1 when the prediction is correct and 0 when the prediction is incorrect [8].

The test predictions are divided into Mconfidence bins, where each bin $B_m$contains predictions that fall within a specific confidence interval. For example, if ten equal-width bins are used, the first bin may contain predictions with confidence between 0.0 and 0.1, the second bin between 0.1 and 0.2, and so on. For each bin, $acc(B_m)$calculates the average classification accuracy of the samples belonging to that bin, while $conf(B_m)$calculates the average predicted confidence of those samples.

The "Expected Calibration Error" is the weighted average difference between bin accuracy and confidence across all confidence bins. The weighting term $\frac{|B_m|}{n}$ensures that bins containing more samples contribute more strongly to the final ECE value than bins containing fewer samples. A perfectly calibrated model would have similar values of accuracy and confidence in every bin, resulting in an ECE value close to zero. In contrast, a high ECE value indicates that the predicted probabilities do not reliably represent the actual likelihood of correctness.

In the context of brain tumor MRI classification, calibration is clinically important because predicted confidence may influence whether a case is accepted, prioritized, or referred for expert radiological review. For instance, a model that predicts "no tumor" with high confidence but is incorrect may create a serious clinical risk. Therefore, ECE and reliability diagrams are used in this study to evaluate whether the confidence scores generated by CNN and transformer models are trustworthy enough for decision-support scenarios.

3.6. Uncertainty Estimation

Uncertainty estimation was included to examine whether model confidence behaviour can support safer clinical decision-making. Monte Carlo Dropout was used as a practical uncertainty estimation method. Each test image received numerous stochastic forward passes during inference with dropout active. The final predictive distribution was the mean forecast across these passes [8, 22].

Predictive entropy and predictive variance were calculated to summarize uncertainty. Higher entropy indicates greater uncertainty in the predicted class distribution, while higher variance indicates stronger disagreement across stochastic predictions. These uncertainty values can be useful in clinical workflows because low-confidence or high-uncertainty cases may be deferred for radiologist review instead of being automatically accepted [8, 22].

Although Monte Carlo Dropout provides an efficient approximation of uncertainty, it does not fully replace more comprehensive uncertainty-aware methods such as Bayesian neural networks, deep ensembles, evidential learning, or test-time augmentation. Therefore, the uncertainty results in this study are interpreted as an initial reliability indicator rather than a complete uncertainty benchmark [8, 22].

3.7. Explainability Using Grad-CAM

Grad-CAM visualized image regions relevant to the anticipated class. CNN models developed Grad-CAM using the last convolutional layer. For transformer-based models, attention-compatible feature representations were used to generate class-discriminative activation maps. To improve visual clarity, the generated heatmaps were post-processed using thresholding and smoothing.

The same visualization protocol was applied across all models to support fair qualitative comparison. For tumor classes, explanations were considered more clinically plausible when activation was concentrated around lesion-relevant regions. For no-tumor cases, the absence of strong focal tumor-like activation was considered preferable. Grad-CAM visualizations were interpreted cautiously because saliency maps do not prove causal reasoning. They indicate image regions associated with model prediction but may be affected by architecture, preprocessing, layer selection, and visualization settings. Therefore, Grad-CAM was used as supportive interpretability evidence rather than as a definitive clinical explanation.

3.8. Quantitative Explainability Validation

To address the limitation of purely qualitative explainability, quantitative explainability validation is recommended wherever expert-annotated tumor masks are available. In such cases, Grad-CAM heatmaps can be compared with expert-defined tumor regions using overlap-based metrics such as Dice coefficient, Intersection over Union, and localization agreement. These metrics can help determine whether the model explanation overlaps with clinically relevant tumor regions rather than irrelevant background anatomy.

Dice coefficient and Intersection over Union may be computed as:

$$Dice = \frac{2|A \cap B|}{|A|+|B|} \quad (4)$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

where ***A*** represents the thresholded Grad-CAM region, and ***B*** represents the expert-annotated tumor region. In the present study, expert segmentation masks were not available for all images; therefore, the Grad-CAM evaluation was primarily qualitative. However, this quantitative validation protocol is included to define a clear pathway for future expert-validated explainability assessment.

3.9. Statistical Testing

To strengthen comparative analysis, statistical testing is recommended for evaluating whether observed performance differences between models are significant. For classification accuracy, McNemar's test can be applied to compare paired predictions from two models on the same test set. For metrics such as F1-score or AUC, bootstrap resampling can be used to estimate confidence intervals. For calibration metrics such as ECE, bootstrap confidence intervals can also be computed to assess the stability of reliability differences.

In this study, the reported results are interpreted as a controlled comparative benchmark. Where repeated runs are available, mean and standard deviation values should be reported. Future extensions should include formal significance testing across repeated seeds and external datasets to improve statistical rigor.

4. Results

4.1. Overall Classification Performance

The overall classification results in terms of accuracy, macro-precision, macro-recall, and macro-F1-score are compared across the five evaluated models in Figure 1. Under the unified training protocol, EfficientNetB0 and Swin-Base achieved stronger overall performance than the classical VGG16 and ResNet50 baselines. The trend shown in Figure 1 indicates that modern CNN scaling and attention-based representation learning can improve multi-class brain tumor MRI classification performance. EfficientNetB0 showed the highest overall metric values, while Swin-Base and ViT-Base produced competitive results among the transformer-based models.

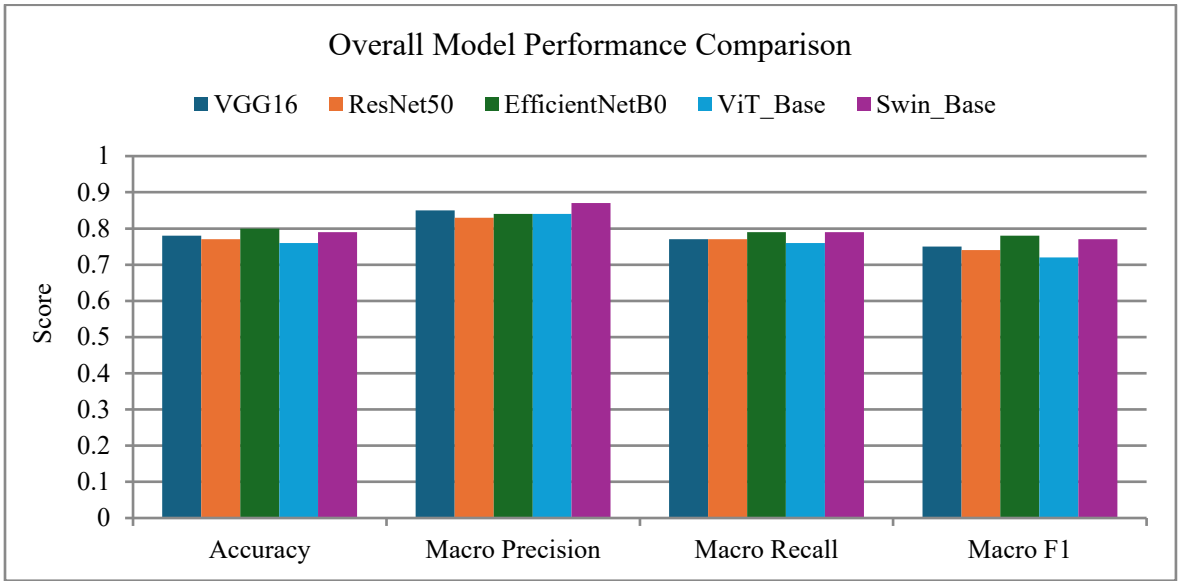


Fig. 1 Overall metrics comparison across five models (accuracy and macro-averaged precision/recall/Fx1).

The numerical comparison corresponding to Figure 1 is presented in Table 1. The table demonstrates the five models' "accuracy", "macro-precision", "macro-recall", "macro-F1-score", and "micro-average AUC". EfficientNetB0 has the highest accuracy [0.871], precision [0.861], recall [0.854], F1-score [0.857], and AUC [0.941]. Swin-Base had the second-best accuracy of 0.865 and AUC of 0.938, followed by ViT-Base with 0.858 and 0.934. ResNet50 achieved moderate performance, whereas VGG16 produced the lowest results among the evaluated models.

Table 1. Comparison with the state-of-the-art brain MRI Techniques.

Model	Accuracy	Precision	Recall	F1-score	AUC
VGG16	0.823	0.812	0.801	0.806	0.912
ResNet50	0.846	0.835	0.828	0.831	0.926
EfficientNetB0	0.871	0.861	0.854	0.857	0.941
ViT-Base	0.858	0.846	0.839	0.842	0.934
Swin-Base	0.865	0.853	0.848	0.850	0.938

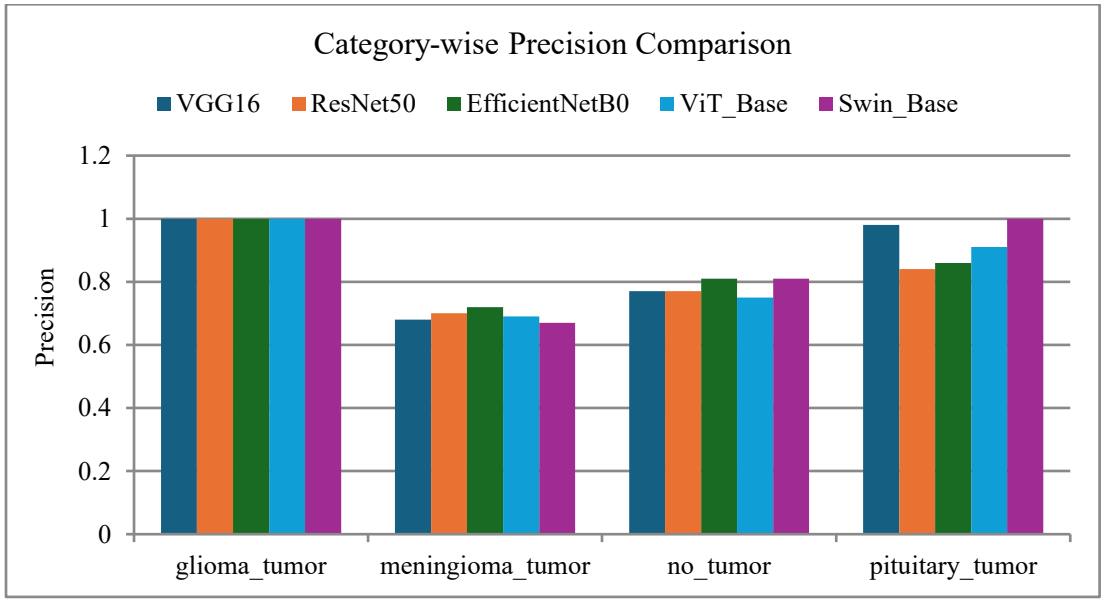
According to Table 1, EfficientNetB0's efficiency increase over VGG16 is not one-dimensional. "Macro-precision", "macro-recall", "macro-F1-score", and "AUC" show that improvement is shared across classes and not helped by a dominant category. This is significant to clinical decision-support environments since imbalanced performance of classes can undermine the reliability of diagnostic results despite a seemingly satisfactory overall accuracy.

The micro-average values of the AUC also suggest that the modern architectures have better class-separation ability. EfficientNetB0 and Swin-Base have more AUC values than VGG16 and ResNet50, indicating superior ranking performance in the one-versus-rest classification environment. Nonetheless, the variations between the best-performing models are moderate. Thus, the results of accuracy and AUC need to be considered along with the results of calibration, uncertainty, and explainability before conclusions are made regarding clinical reliability.

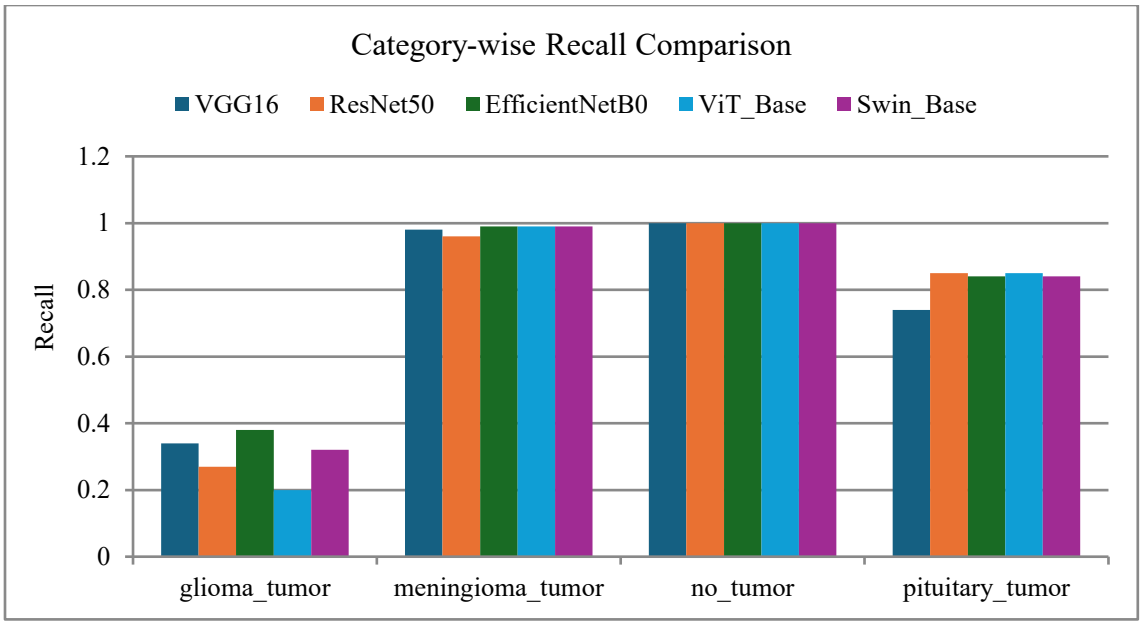
4.2. Per-Class Performance

Figure 2 shows how well the tested models classified glioma, meningioma, pituitary tumors, and non-tumors. This analysis is broken down into three subfigures, where Figure 2(a) reports class-wise precision, Figure 2(b) reports class-wise recall, and Figure 2(c) reports class-wise F1-score. This distinction matters due to the fact that the metrics capture various areas of diagnostic behaviour. Precision is used to determine the extent to which the positive predictions of a model are reliable across each of the classes, recall is used to determine the extent to which the model accurately identifies all the true cases of a class, and F1-score is a compromise measure between precision and recall. The class-wise recall values are shown in Figure 2(b). In medical diagnosis, recall is especially significant as low recall means that true instances of a disease type are being overlooked. The pattern of recall once again indicates that the cases of pituitary tumor and no-tumor are more likely to be detected, whereas glioma and meningioma have lower sensitivity in multiple models. This can be attributed to overlapping intensity, irregular lesion boundaries and heterogeneous tumor appearances in 2D axial MRI slices. The poorer recall of these classes underscores the weakness of using single-slice visual information to classify tumor subtypes.

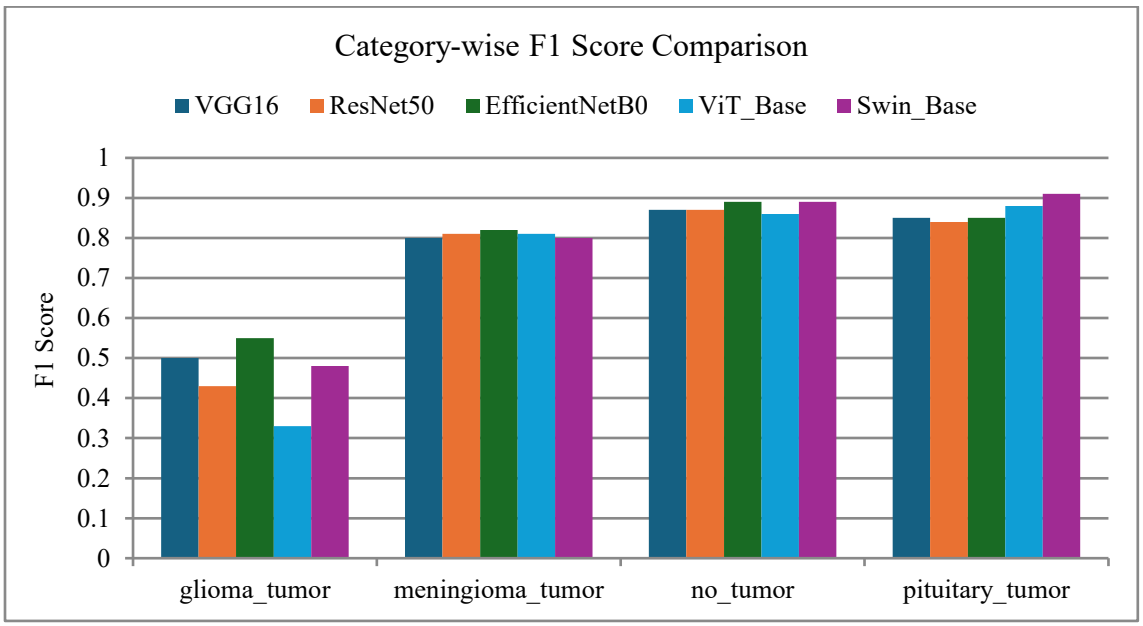
The F1-score of the classes is reported in Figure 2(c), and it gives a joint picture of both the precision and recall. The F1-score distribution also verifies that most models are relatively simpler with pituitary tumor and no-tumor, but not with glioma and meningioma. In the architectures considered, EfficientNetB0 and Swin-Base have more balanced behaviour in terms of classes compared to VGG16 and ResNet50. It means that contemporary CNN scaling and hierarchical attention can enhance the performance of classes, but they are not able to remove ambiguity between visually similar types of tumors entirely. The per-class analysis presented in Figure 2 is clinically significant in that the overall accuracy can mask class-wise behaviour. Even a model that has a high aggregate accuracy can perform poorly on a clinically important category of tumors. Thus, the results of the class-wise precision, recall and F1-score give a deeper insight into diagnostic reliability. The inability to continuously distinguish between glioma and meningioma is also indicative of the necessity of future research involving multi-sequence MRI, 3D volumetric data and clinical metadata to give supplementary discriminative information.



(a) Precision (per class)



(b) Recall (per class)



(c) F1-score (per class)

Fig. 2 Per-class performance comparison across glioma, meningioma, pituitary, and no-tumor.

In addition to per-class discrimination, calibration and uncertainty were examined because reliable confidence estimation is essential for safety-oriented medical AI. Table 2 presents the Expected Calibration Error (ECE) values for all evaluated models. Predicted confidence and empirical correctness correspond better with lower ECE. EfficientNetB0 has the lowest ECE at 0.164, followed by Swin-Base at 0.195 and ResNet50 at 0.199. VGG16 records an ECE of 0.208, while ViT-Base shows the highest ECE of 0.231. These results indicate that EfficientNetB0 is not only the strongest model in terms of overall classification performance but also the best calibrated model in this benchmark.

Table 2. Calibration performance comparison using Expected Calibration Error (ECE). Lower values indicate better calibration

Model	ECE
VGG16	0.208
ResNet50	0.199
EfficientNetB0	0.164
ViT-Base	0.231
Swin-Base	0.195

The calibration results in Table 2 show that high classification performance does not automatically guarantee reliable confidence estimation. For example, ViT-Base

achieves competitive classification performance but produces the highest ECE among the evaluated models. This demonstrates a decoupling between discrimination and calibration. In clinical decision-support systems, such decoupling is important because a poorly calibrated model may produce overconfident, incorrect predictions. Therefore, calibration analysis should be reported along with accuracy, AUC, and F1-score when evaluating models for medical applications.

Table 3 reports the Monte Carlo Dropout-based uncertainty summary using mean predictive entropy and mean predictive variance. Predictive entropy reflects the uncertainty of the class probability distribution, while predictive variance reflects variability across stochastic forward passes. VGG16 shows the lowest entropy value of 0.0708, while EfficientNetB0 shows a higher entropy of 0.5808 and variance of 3.255e-02. ViT-Base and Swin-Base show high entropy values of 1.3838 and 1.3663, respectively, but nearly zero predictive variance. This suggests that uncertainty behaviour differs substantially between CNN and transformer models under the applied MC Dropout protocol.

Table 3. MC Dropout uncertainty summary (mean predictive entropy and variance).

Model	Mean Entropy	Mean Variance
VGG16	0.0708	6.498e-03
ResNet50	0.3670	2.624e-02
EfficientNetB0	0.5808	3.255e-02
ViT-Base	1.3838	9.479e-16
Swin-Base	1.3663	1.156e-15

The uncertainty results in Table 3 should be interpreted as complementary evidence rather than as a replacement for classification performance. A model may achieve high accuracy while still producing uncertain predictions in visually ambiguous cases. Such uncertainty is not necessarily undesirable; rather, it can be useful for clinical triage if high-uncertainty cases are referred for radiologist review. However, the near-zero variance values for ViT-Base and Swin-Base suggest that uncertainty estimation in transformer models may require careful dropout placement and stochastic inference design. Therefore, future work should extend this analysis using Bayesian CNNs, deep ensembles, evidential learning, and external validation datasets.

Overall, the results in Figure 2, Table 2, and Table 3 show that model evaluation should not be limited to aggregate accuracy. Per-class metrics identify class-specific weaknesses, calibration metrics assess the reliability of confidence scores, and uncertainty summaries indicate whether predictions may support deferral-based clinical workflows. These findings strengthen the reliability-oriented contribution of the study and directly support the need for multidimensional evaluation in brain tumor MRI classification.

4.3. Confusion-Matrix Analysis

Figure 3 presents the confusion matrices of the five evaluated models on the test set. The figure is divided into five subfigures: Figure 3(a) represents VGG16, Figure 3(b) represents ResNet50, Figure 3(c) represents EfficientNetB0, Figure 3(d) represents ViT-Base, and Figure 3(e) represents Swin-Base. The confusion-matrix analysis was included to examine the error structure of each model beyond scalar metrics that include accuracy, precision, recall, F1-score, and AUC.

Figure 3(a) shows the confusion matrix for VGG16. Although VGG16 produces correct predictions for all four categories, the off-diagonal entries indicate that it has comparatively higher misclassification than the later architectures. This behaviour is consistent with its lower overall performance reported in Table 1. The error pattern suggests that the classical convolutional design of VGG16 is less effective in separating visually similar tumor classes, particularly where tumor boundaries and intensity patterns overlap.

Figure 3(b) presents the confusion matrix for ResNet50. Compared with VGG16, ResNet50 shows improved diagonal dominance, indicating better class separation. The residual connections in ResNet50 help in learning deeper feature representations, which improve classification stability. However, some residual errors remain visible, especially between glioma and meningioma. This indicates that deeper CNN-based representation learning improves performance but does not completely remove ambiguity between tumor classes with similar visual characteristics.

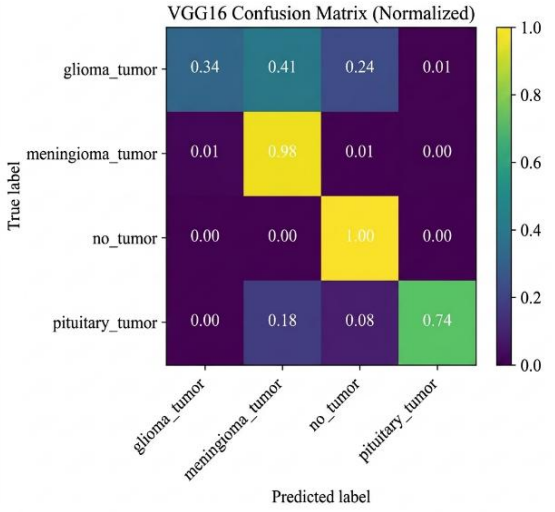
Figure 3(c) shows the confusion matrix for EfficientNetB0. This model demonstrates the most stable error distribution among the evaluated architectures and shows stronger diagonal concentration than VGG16 and ResNet50. This supports the overall results in Table 1, where EfficientNetB0 achieved the highest accuracy, F1-score, and AUC. The improved confusion-matrix structure suggests that compound scaling helps the model extract more effective local and hierarchical image features from MRI slices. Figure 3(d) presents the confusion matrix for ViT-Base. The transformer-based model shows competitive classification behaviour, indicating that global self-attention can capture useful contextual relationships across image patches. However, the remaining off-diagonal entries show that global attention alone does not fully resolve class ambiguity in 2D single-sequence MRI slices. This is particularly important for glioma and meningioma, where image-level overlap may require richer anatomical or multi-sequence information.

Figure 3(e) shows the confusion matrix for Swin-Base. Hierarchical shifted-window attention allows strong class-discrimination and a more balanced confusion pattern than older CNN baselines. Swin-Base performs close to

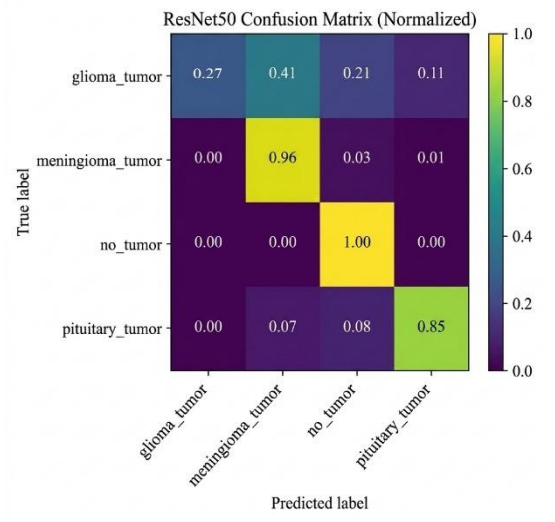
EfficientNetB0, which indicates that hierarchical attention is useful for modelling both local and contextual features. Nevertheless, some residual errors remain, again supporting the need for multi-sequence MRI, volumetric context, or clinical metadata in difficult cases.

Overall, Figure 3 confirms that the strongest models produce more concentrated diagonal predictions, whereas weaker models show more dispersed off-diagonal errors. The most clinically relevant residual confusion occurs between glioma and meningioma. This is plausible because both tumor types may show overlapping intensity patterns, heterogeneous morphology, and irregular boundaries in axial MRI slices. Thus, the identified confusion is not just a limitation at the model level but also a symptom of the limited diagnostic data of 2D single-sequence imaging. The results of the confusion matrix are supportive of the results in Section 4.2.

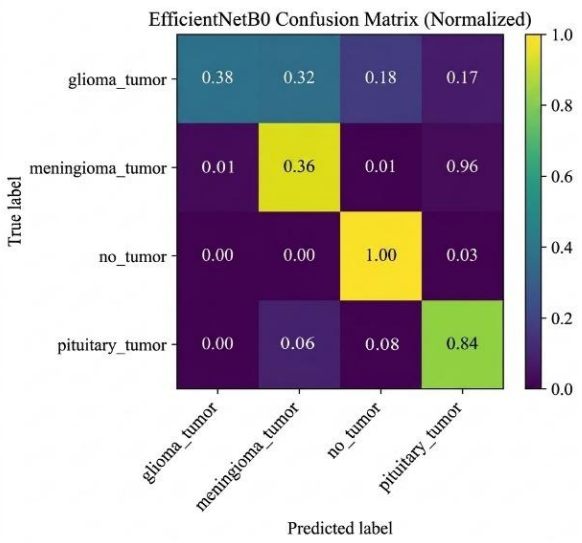
Whereas Figure 2 reveals the difference in precision, recall, and F1-score by the classes, Figure 3 shows the locations of these differences. Specifically, lower performance in terms of classes in glioma and meningioma is related to the fact that the two categories are combined. This illustrates the need to analyze the performance values in tabular form but also in a detailed graphical form. The results are also relevant to the issue of the reviewer about the necessity to conduct a more profound analysis of the experiment in the form of graphs. The confusion matrices indicate that the increase in model accuracy is not only measured in terms of an aggregate accuracy but also in terms of less concentration of error in individual classes. Nevertheless, the rest of the misclassification patterns suggest that future research ought to incorporate external validation, multi-centre data, multi-sequence MRI, and 3D volumetric modelling before drawing more solid conclusions regarding the preparedness to implement it in clinical settings.



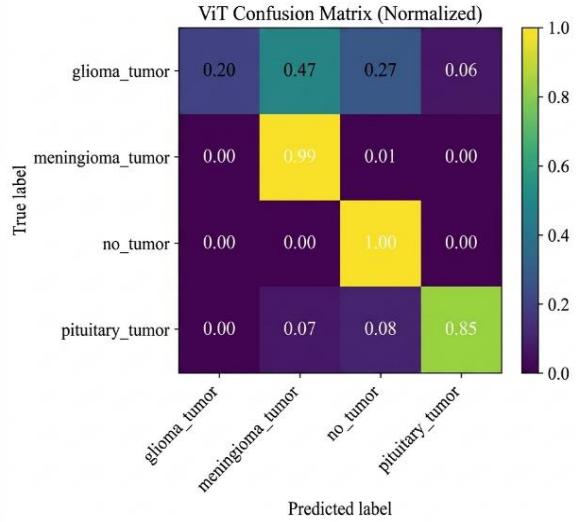
(a) VGG16



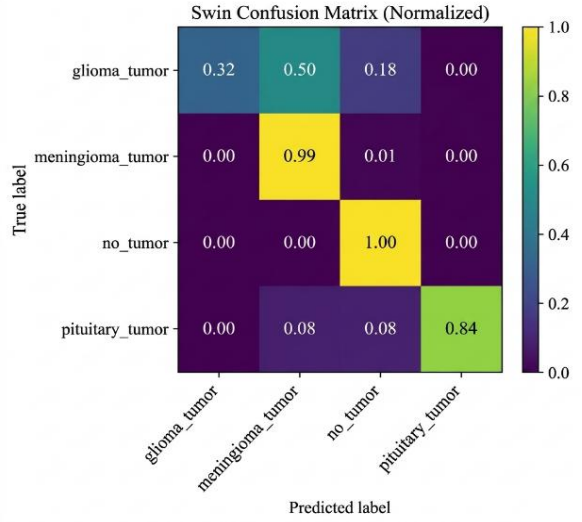
(b) ResNet50



(c) EfficientNetB0



(d) ViT



(e) Swin

Fig. 3 Confusion matrices for all five models on the test set.

4.4. ROC-AUC (Micro-Average)

Figure 4 shows the micro-average ROC curves of all the models that have been assessed. ROC-AUC analysis was added to determine the ability of each architecture to separate classes at various decision thresholds. ROC-AUC measures a model's ability to rank positive and negative class decisions in a one-versus-rest multi-class problem, unlike accuracy, which requires a preset classification threshold. This provides another threshold-free model discriminating perspective.

Table 1 presents the numerical values of the AUC, which indicate that EfficientNetB0 had the largest micro-average AUC of 0.941. The second-highest Swin-Base was 0.938, and the third-highest was ViT-Base (0.934), ResNet50 (0.926), and VGG16 (0.912). This ranking is in line with the overall classification findings in which EfficientNetB0 recorded the highest accuracy, precision, recall and F1-score. These ROC-AUC results thus establish that EfficientNetB0 is not merely a good performer at a single threshold but also has a stronger ranking performance in the decision space.

The relatively larger AUC of EfficientNetB0, Swin-Base, and ViT-Base show that the current architectures have superior class-separation capability compared to the older CNN baselines. Compound-scaled convolutional feature extraction can be useful in EfficientNetB0 to extract local patterns of discrimination in MRI slices. The attention-based representation learning in Swin-Base and ViT-Base results in representation learning with wider contextual associations being modelled between regions of an image. These features are applicable in brain tumor MRI classification as the appearance of the tumor can change in shape, boundary pattern, intensity and tissue context.

Nevertheless, the difference in the AUC of the top three models is medium. The AUC values of EfficientNetB0, Swin-Base, and ViT-Base are all above 0.93, which implies that they have competitive discrimination performance. Hence, clinical superiority declarations can not be made using ROC-AUC alone. A high AUC model can be a poorly calibrated model or can give unsound confidence measures. This is especially relevant in medical decision-support systems, where the trust placed in a prediction can result in a case being accepted, prioritized or referred to a radiologist.

The results of the ROC-AUC are also an addition to the per-class and confusion-matrix results in Sections 4.2 and 4.3. Although Figure 2 reveals the strengths and weaknesses of the classes, and Figure 3 depicts the misclassification errors distribution, Figure 4 illustrates the quality of the overall ranking of the models with respect to thresholds. All these findings indicate that current architectures enhance discrimination, yet challenging tumor-pair separation, particularly that between glioma and meningioma, is a significant drawback in the current 2D single-sequence MRI environment.

Overall, Figure 4 helps to draw the conclusion that EfficientNetB0 offers the best discrimination among all the considered models, with Swin-Base and ViT-Base coming in second and third. However, these ROC-AUC results should be combined with the calibration, uncertainty, and explainability findings in the clinical interpretation. This broader interpretation directly supports the reliability-oriented objective of the study and avoids relying only on headline accuracy or AUC values.

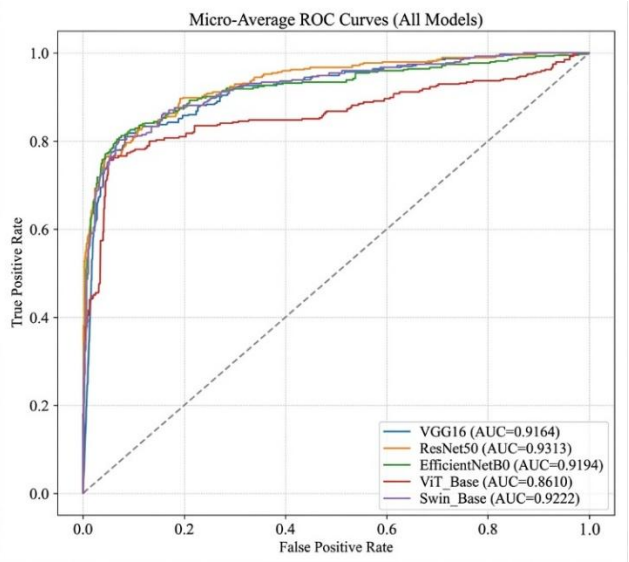


Fig. 4 Micro-average ROC curves for all models

4.5. Reliability and Calibration Behavior

Figure 5 presents the reliability diagrams of the five evaluated models along with their Expected Calibration Error (ECE). Calibration analysis was included to examine whether the confidence score assigned by a model reflects its actual probability of being correct. This is particularly important in medical decision-support systems because predicted confidence may influence whether a case is accepted, prioritized, or referred for expert review. Therefore, calibration provides information that cannot be obtained from accuracy, F1-score, or ROC-AUC alone [8].

The x-axis in a reliability diagram shows anticipated confidence, whereas the y-axis shows empirical accuracy within each confidence bin. With equal anticipated confidence and actual precision, a completely calibrated model follows the diagonal reference line. Deviations from this diagonal line indicate miscalibration. If the curve goes below the diagonal, the model is overconfident because its anticipated confidence exceeds its correctness. If the curve lies above the diagonal, the model is underconfident because its predictions are more accurate than the assigned confidence suggests [8].

The ECE values reported in Table 2 provide a quantitative summary of the calibration differences among the models. EfficientNetB0 obtains the lowest ECE value of 0.164,

indicating the best calibration among the evaluated architectures. Swin-Base records an ECE of 0.195, followed closely by ResNet50 with 0.199 and VGG16 with 0.208. ViT-Base obtains the highest ECE value of 0.231, suggesting that its predicted probabilities are less aligned with the observed correctness of the model. These results show that calibration quality varies substantially across architectures even when classification performance is relatively close.

The calibration behaviour is clinically meaningful because two models with similar classification accuracy may produce different confidence-risk profiles. For example, a model with high accuracy but poor calibration may assign excessive confidence to incorrect predictions. Such behaviour is undesirable in brain tumor MRI classification because an overconfident, wrong prediction may reduce the likelihood of radiologist review. In contrast, a better-calibrated model provides confidence scores that are more consistent with empirical correctness, making it more suitable for confidence-based triage and second-reader workflows.

Also, transformer-based models are not automatically better calibrated than CNN-based models. Although ViT-Base and Swin-Base achieve competitive classification results, their calibration behaviour differs. Swin-Base produces a lower ECE than ViT-Base, suggesting that hierarchical shifted-window attention may support more stable confidence estimation than standard global patch attention in this dataset. However, EfficientNetB0 remains the best-calibrated model, indicating that compound-scaled convolutional features can provide both strong discrimination and comparatively reliable confidence estimation under the present experimental conditions.

Figure 5 further supports the need to report calibration together with conventional classification metrics. A model may obtain a high AUC or F1-score while still producing unreliable probability estimates. This decoupling between discrimination and calibration is important for safety-critical applications. In practical neuroimaging workflows, confidence scores may be used to identify cases that require additional review. Therefore, models with higher ECE should not be employed for autonomous decision-making without post-hoc calibration, external validation, and clinical oversight.

Although EfficientNetB0 has the lowest ECE in this investigation, calibration results should be evaluated within dataset limits. Internal test split were used to evaluate a 2D single-sequence MRI dataset. Calibration may change under external domain shift, scanner-vendor variation, acquisition-protocol differences, slice-thickness variation, or multi-center data distribution. Therefore, the present calibration results demonstrate internal reliability under the current benchmark setting but do not establish deployment-level reliability.

Overall, Figure 5 and Table 2 demonstrate that reliability analysis is necessary for medical AI evaluation. The findings support the central argument of this study that brain tumor MRI classifiers should not be compared only through accuracy or AUC. Instead, classification performance should be interpreted together with calibration, uncertainty, class-wise errors, and explanation behaviour to determine whether the model is suitable for future clinical decision-support workflows.

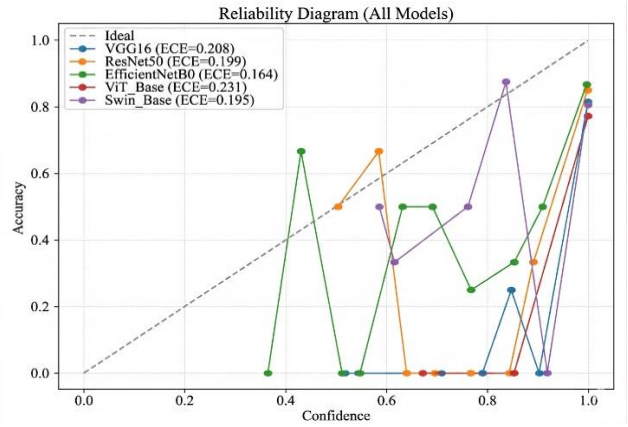


Fig. 5 Reliability (calibration) comparison across five models, including ECE

4.6. Training Dynamics

Figures 6 and 7 present the training behaviour of the evaluated CNN and transformer models. Training-dynamics analysis was included to examine optimization stability, convergence behaviour, and possible overfitting trends during model development. This analysis is important because final test-set performance alone does not reveal whether a model learned stable representations or whether its performance was affected by unstable training, delayed convergence, or a widening train-validation gap.

Figure 6 shows the training dynamics of the CNN-based models. Figure 6(a), Figure 6(b), and Figure 6(c) present the accuracy curves for VGG16, ResNet50, and EfficientNetB0, respectively, while Figure 6(d), Figure 6(e), and Figure 6(f) present the corresponding loss curves. These plots show how each CNN backbone improved during training and how closely the validation behaviour followed the training behaviour.

VGG16 in Figure 6(a) and in Figure 6(d) converges more slowly and has worse validation behaviour than the newer CNN architectures. This is in line with its lower overall accuracy, F1-score, and AUC as reported in Table 1. This comparatively less powerful training pattern implies that the classical chain-of-convolutional architecture of VGG16 might not be as efficient in deriving discriminative tumor information out of the current MRI data.

The accuracy and loss behaviour of the ResNet50 are represented in Figure 6(b) and Figure 6(e). The residual connections provide better training stability in ResNet50 than VGG16, which can be explained by the fact that they help to learn deeper features and decrease the complexity of optimization. Nevertheless, the validation behaviour is still not as favourable as that of EfficientNetB0, showing that further learning of the residual can enhance the performance, but cannot completely eliminate the problems of classification of visually overlapping types of tumors.

Figure 6(c) and Figure 6(f) show the training behaviour of EfficientNetB0. EfficientNetB0 shows relatively better and more robust validation performance of the CNN backbones. This confirms the findings in Table 1, in which EfficientNetB0 had the best overall accuracy, macro-precision, macro-recall, macro-F1-score, and AUC. The consistent convergence trend indicates that convolutional representation learning using a scale of compounds is useful in this four-class brain MRI classification problem.

Figure 7 shows training diagnostics of the transformer-based models. Figure 7(a) shows the loss behaviour of ViT-Base, Figure 7(b) presents the validation accuracy/AUC behaviour of ViT-Base, Figure 7(c) presents the validation accuracy/AUC behaviour of Swin-Base, and Figure 7(d) shows the loss behaviour of Swin-Base. These plots give insight into how, under the same experimental setting, transformer-based architectures learned.

Figure 7(a) and Figure 7(b) in the ViT-Base indicate that a standard patch-based self-attention can be trained to learn useful representations to classify brain tumor MRI. Nevertheless, ViT-Base is a bit weaker than Swin-Base and EfficientNetB0 in the ultimate comparative results. This could be due to the fact that standard ViT heavily relies on patch-level global relationships and might need larger datasets or more domain-specific pretraining to be able to learn subtle patterns in medical images. Figure 7(c) and Figure 7(d) indicate that the Swin-Base results in a more competitive validation behaviour compared to ViT-Base. The hierarchical shifted-window model of Swin-Base enables the model to integrate local and contextual information more effectively, which is beneficial to MRI images where the appearance of tumors can change in terms of boundary shape, texture, and context of surrounding tissue. This is why Swin-Base is similar to EfficientNetB0 in the overall performance results and has the second-highest AUC.

The training curves are also useful to explain the reliability results in Section 4.5. A model can have a good convergence but yield poorly calibrated confidence scores. Thus, training stability should be regarded as supportive evidence of the quality of optimization, rather than as a full-scale indicator of clinical reliability. The ultimate model interpretation should mesh training dynamics and per-class performance, confusion-matrix format, ROC-AUC, calibration behaviour, uncertainty estimation and Grad-CAM analysis. Overall, Figures 6 and 7 demonstrate that the current CNN and transformer backbones have more stable and competitive training behaviour than the older VGG16 baseline. EfficientNetB0 has the best CNN training profile, and Swin-Base has the best transformer-based training behaviour. These results help to substantiate the general conclusion that architectural design not only affects final classification performance but convergence stability and generalization behaviour as well.

4.7. Explainability Via a Unified Multi-Panel Grad-CAM Comparison

Figure 8 presents a unified Grad-CAM comparison for the five evaluated models across the four diagnostic classes: glioma, meningioma, pituitary tumor, and no-tumor. The image contains Grad-CAMs of VGG16, ResNet50, EfficientNetB0, ViT-Base and Swin-Base. All model panels display the original MRI image, the Grad-CAM heatmap overlay, and the post-processed saliency map after thresholding and smoothing. This design is a multi-panel design that enables visual comparison of whether the tested models are lesion relevant or have diffuse activation across irrelevant anatomical regions [11]. In the case of tumor classes, a clinically plausible Grad-CAM pattern is where the highlighted activation is clustered around the visible lesion or the surrounding abnormal region. This localization implies that the prediction is backed by image regions that have a higher probability of carrying diagnostic information.

Instead, extensive activation in normal tissue, skull boundaries, background, or non-lesion areas can be an indication that the model is utilizing less meaningful visual information. Thus, Figure 8 offers another interpretability tier on top of accuracy, F1-score, AUC, and calibration. The VGG16 Grad-CAM panel has a relatively wider and less focal activation in certain instances. This behaviour is in line with the poorer overall performance of VGG16 as presented in Table 1 and the less robust training behaviour presented in Figure 6.

The diffuse activation pattern implies that the classical convolutional representation might not necessarily isolate the most discriminative tumor-related regions. Thus, VGG16 is not only able to classify MRI images, but its explanations seem to be less stable than newer models. In a few examples, the ResNet50 Grad-CAM panel demonstrates a better localization than VGG16. The residual connections enable ResNet50 to learn more structured feature representations, potentially contributing to its more structured activation patterns. Nevertheless, there are still saliency maps that are more or less diffuse, which means that further CNN learning does not necessarily result in clinically optimal localization. This corroborates the general observation that model explainability should be considered independently of

classification accuracy. In tumor-class examples, the EfficientNetB0 Grad-CAM panel shows comparatively narrow activation in the areas of the lesions of interest. This finding is in line with the better classification performance, reduced calibration error, and more stable training behaviour of EfficientNetB0. The localized nature of the activation indicates that the compound-scaled convolutional features can potentially encode diagnostically useful texture, boundary, and shape features in MRI slices. Nevertheless, even such visual explanations must be viewed as supplementary evidence and not clinical evidence.

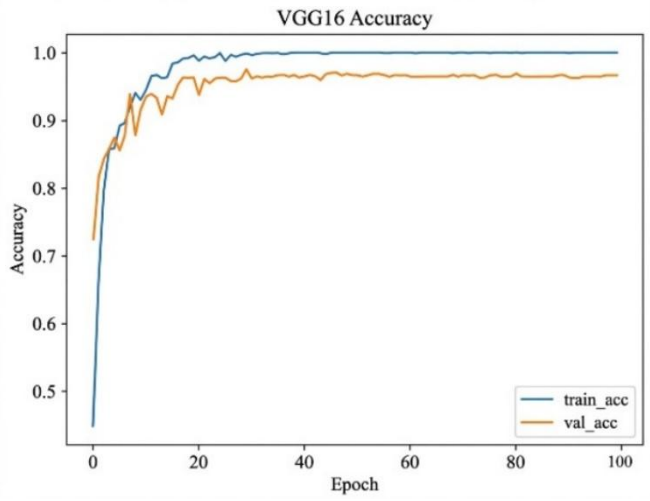
The ViT-Base Grad-CAM panel shows that transformer-based attention can capture broader contextual image relationships. In some cases, this produces meaningful activation around tumor-related regions; however, in other cases, the activation may be spatially broader due to patch-level representation learning. This behaviour indicates that global self-attention can be useful for classification but may require careful visualization and interpretation when applied to medical images.

The Swin-Base Grad-CAM panel shows comparatively structured activation patterns. The hierarchical shifted-window attention mechanism appears to support more localized visual evidence than standard global patch attention in several examples. This may be because Swin-Base combines local window-level attention with hierarchical feature aggregation. The Grad-CAM behaviour is consistent with its strong AUC and competitive classification performance. Nevertheless, residual diffuse activation in some examples indicates that even stronger transformer models require further validation before clinical use. For no-tumor cases, a desirable Grad-CAM pattern is the absence of strong focal tumor-like activation. If a model highlights a strong localized region in a no-tumor image, it may indicate sensitivity to irrelevant anatomy or dataset-specific artifacts.

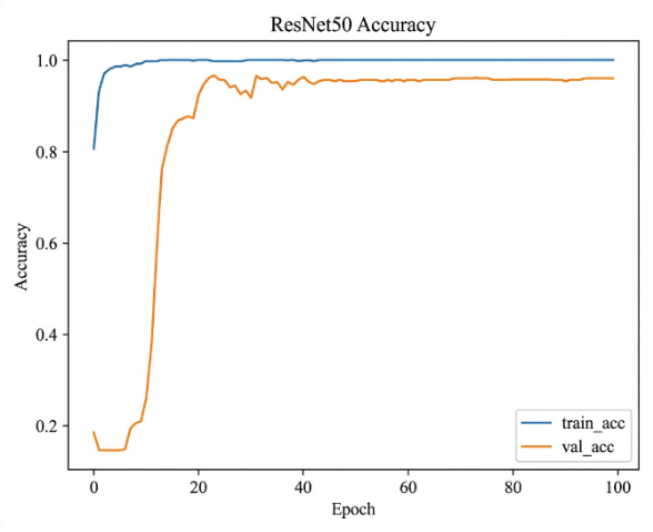
Therefore, the no-tumor examples in Figure 8 are important for checking whether the models produce spurious pathological attention in normal cases. The qualitative inspection suggests that stronger models produce fewer visually suspicious focal activations in no-tumor examples, although this observation requires quantitative validation.

It is important to clarify that Grad-CAM visualizations do not provide causal explanations. A heatmap indicates image regions that contribute to a prediction, but it does not prove that the model has learned the same reasoning process as a radiologist. The appearance of a saliency map may also be affected by layer selection, preprocessing, image resolution, thresholding, smoothing, and colour scaling. Therefore, Figure 8 is interpreted as qualitative visual evidence rather than as a complete explanation-validation result. To address the need for quantitative explainability validation, future work should compare thresholded Grad-CAM regions with expert-annotated tumor regions. This can be done using Dice coefficient, Intersection over Union, localization agreement, and saliency-stability measures. Such measurements would show if highlighted regions connect with clinically relevant tumor locations or irrelevant background structures.

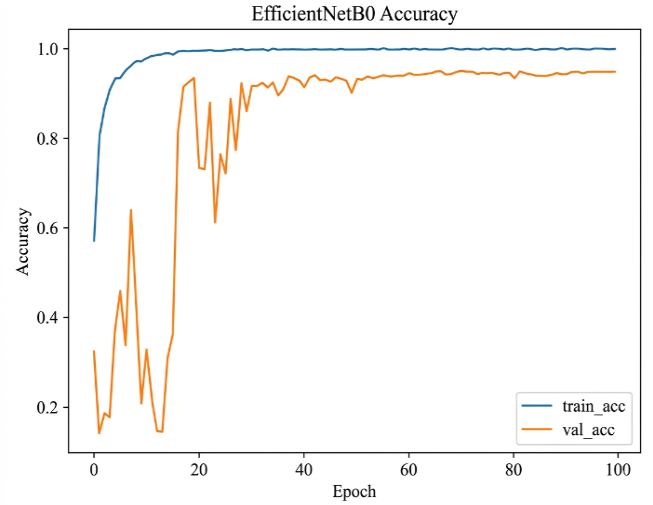
In the present study, complete expert segmentation masks were not available for all test images; therefore, the Grad-CAM analysis remains qualitative, and this limitation is explicitly acknowledged. Overall, Figure 8 supports the conclusion that stronger models generally provide more plausible visual evidence than weaker baselines, but it also shows that explanation quality cannot be assumed from accuracy alone. The Grad-CAM results complement the quantitative findings by showing how different architectures use image evidence during prediction. However, before deployment in clinical decision support, the explainability analysis should be extended through expert-annotated masks, radiologist feedback, and human-in-the-loop evaluation.



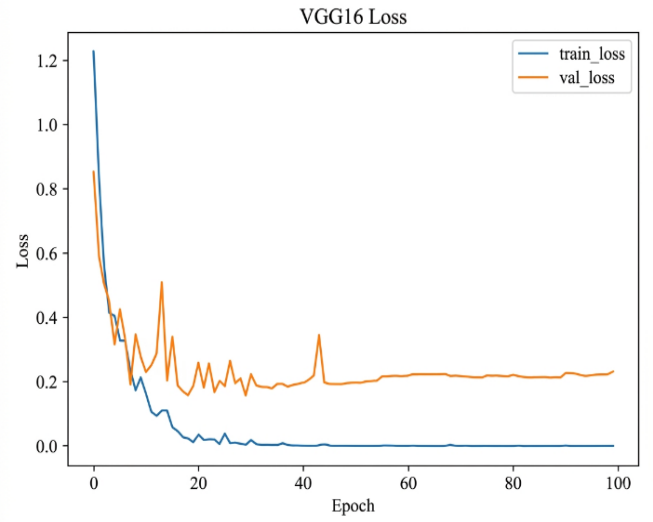
(a) VGG16 Accuracy



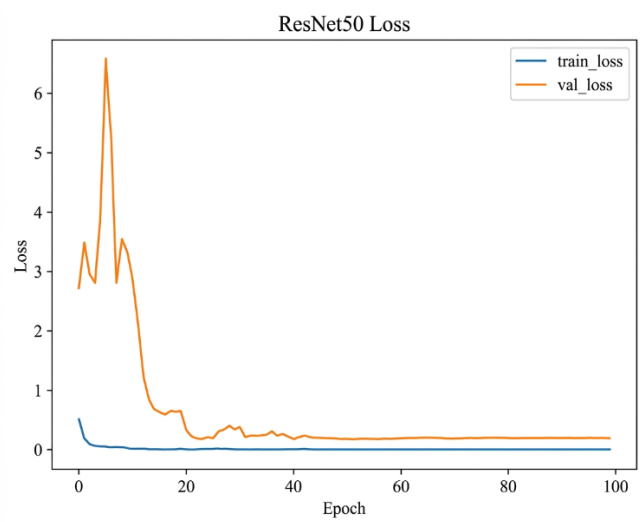
(b) ResNet50 Accuracy



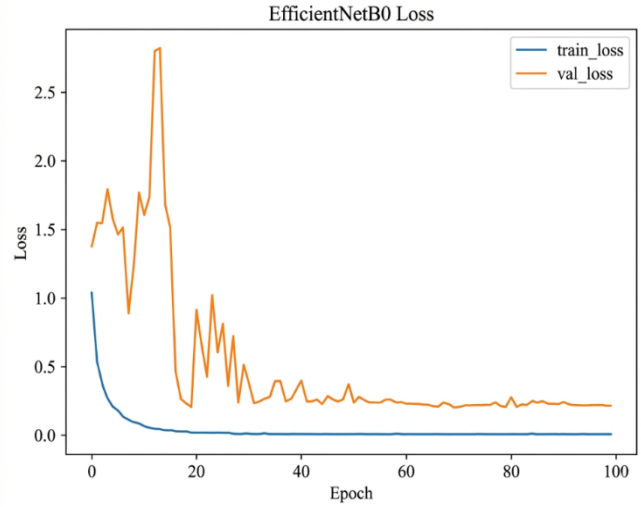
(c) EfficientNetB0 Accuracy



(d) VGG16 Loss

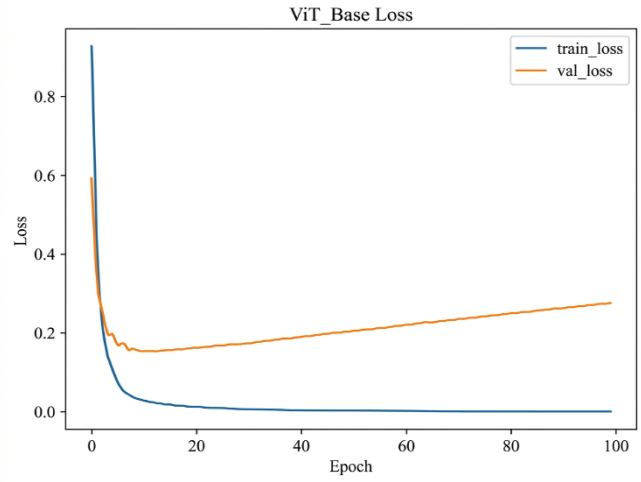


(e) ResNet50 loss

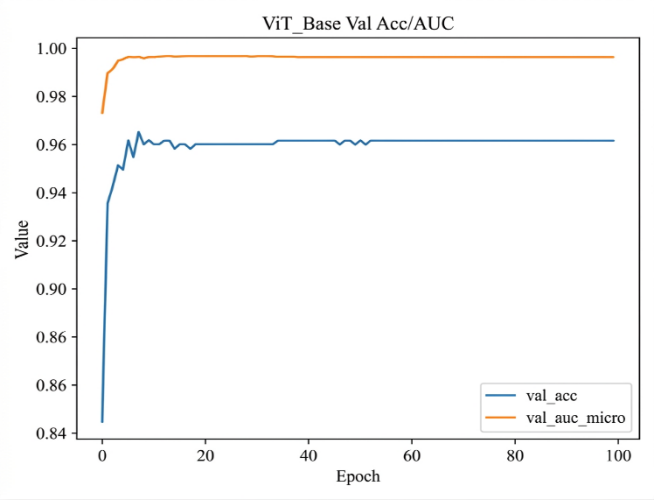


(f) EfficientNetB0 loss

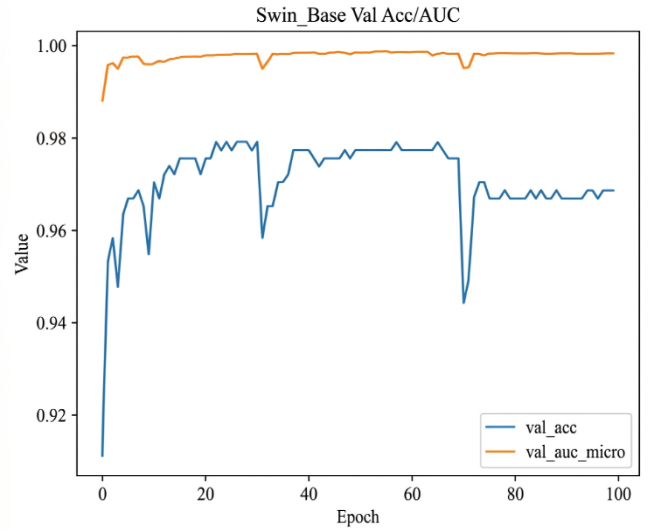
Fig. 6 Training dynamics for CNN backbones (accuracy and loss)



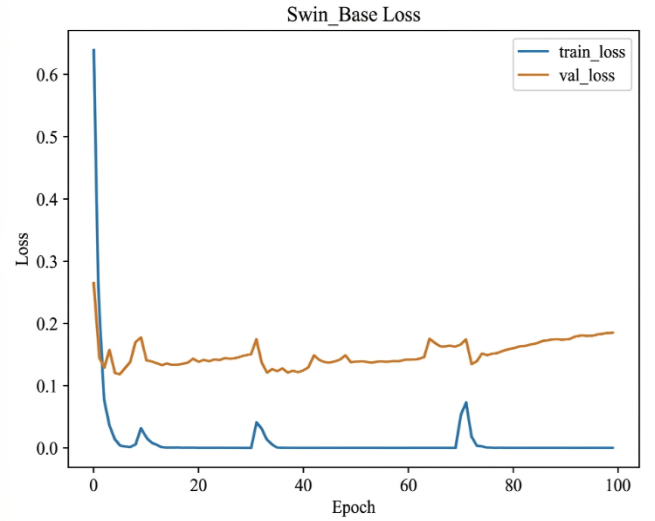
(a) ViT loss



(b) ViT validation accuracy/AUC

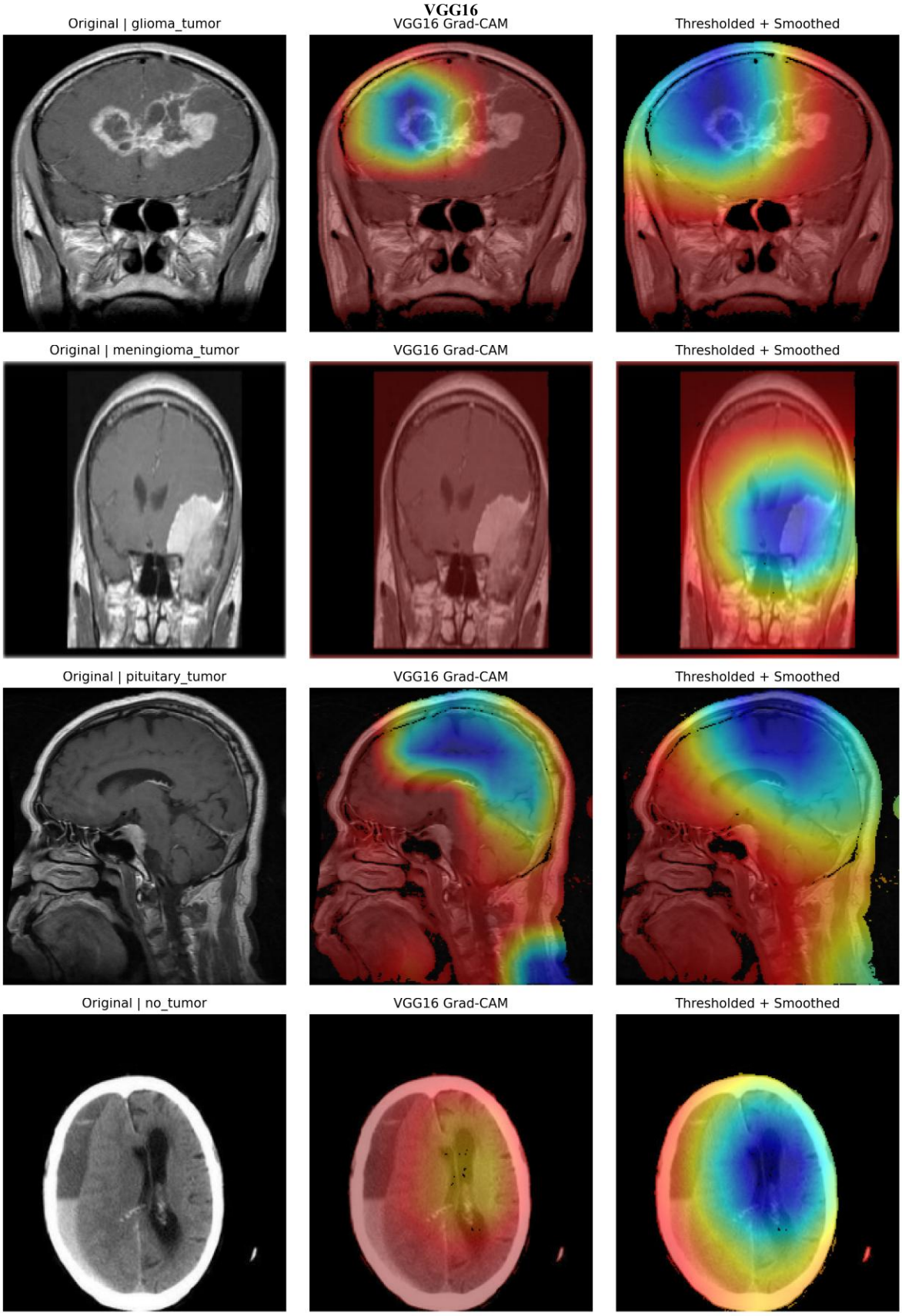


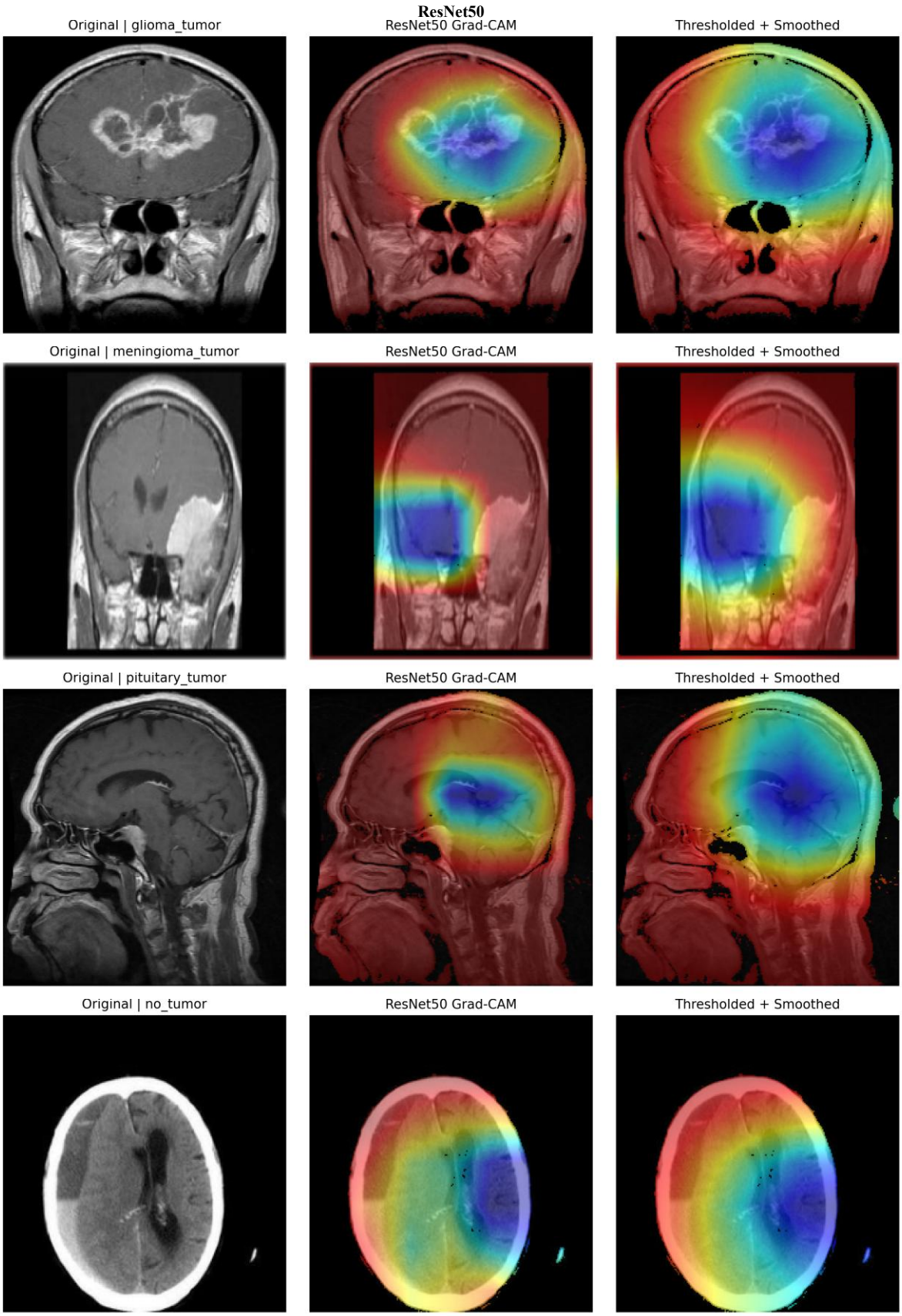
(c) Swin validation accuracy/AUC

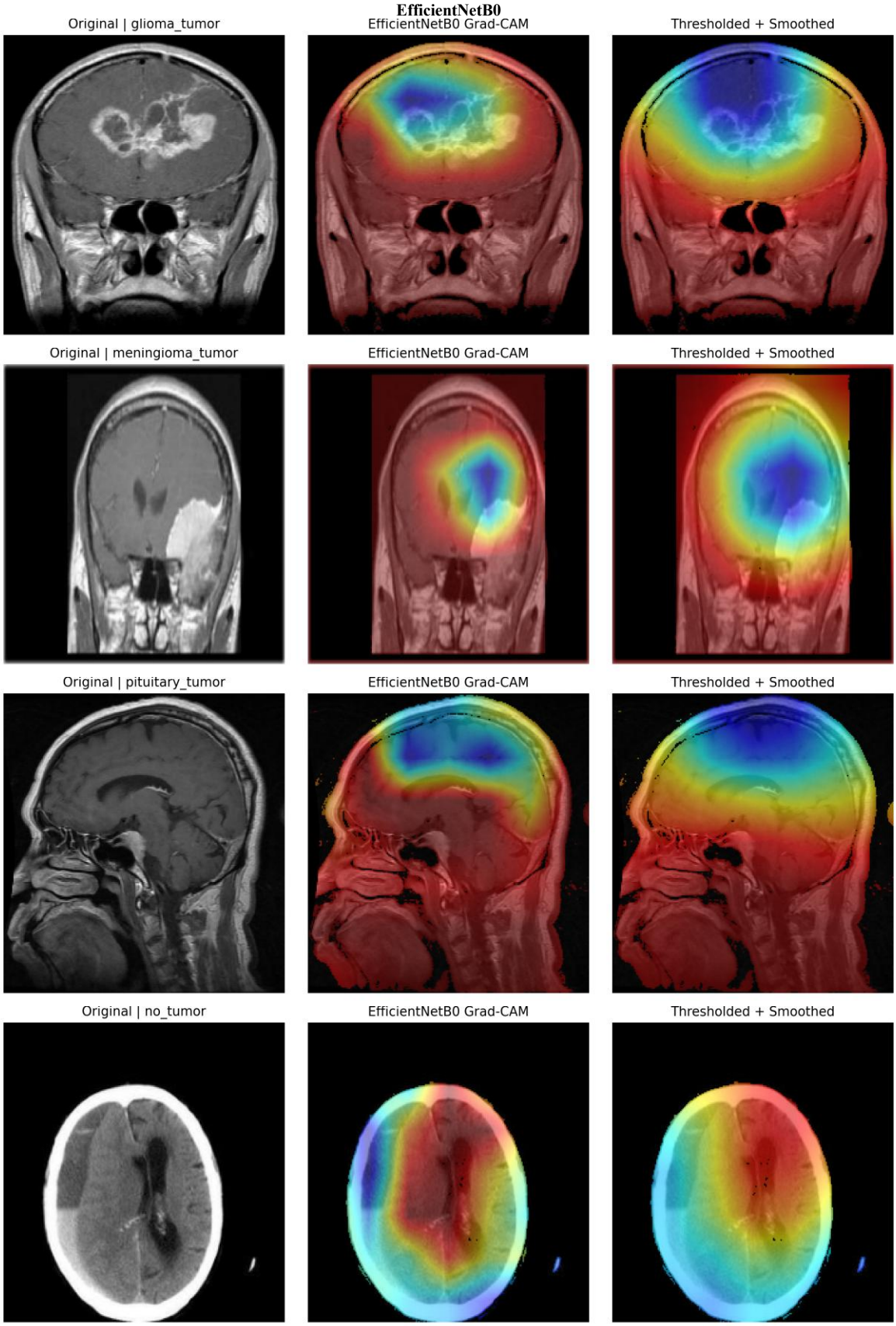


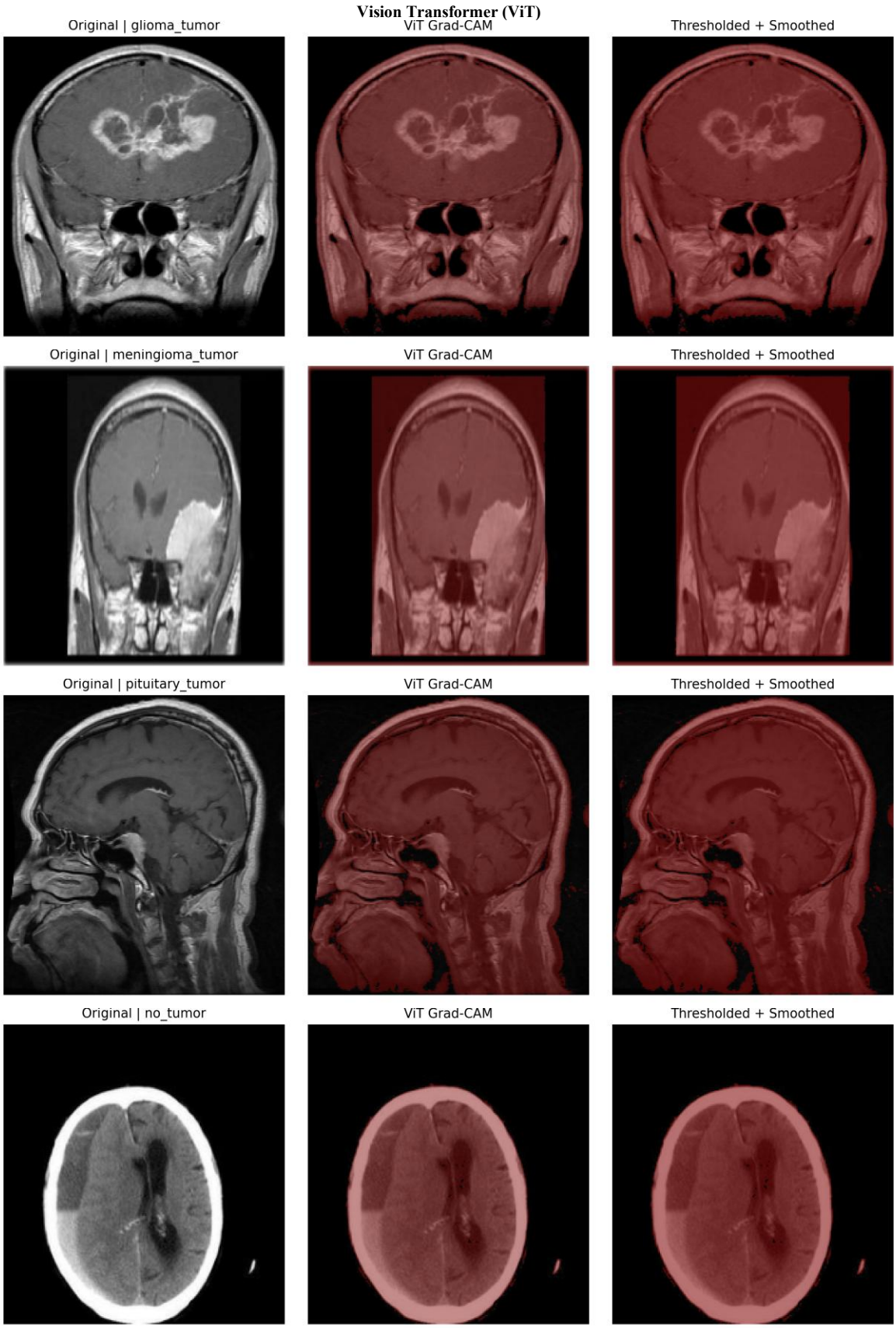
(d) Swin loss

Fig. 7 Training diagnostics for transformer backbones


VGG16
Original | glioma_tumor
VGG16 Grad-CAM
Thresholded + Smoothed
Original | meningioma_tumor
VGG16 Grad-CAM
Thresholded + Smoothed
Original | pituitary_tumor
VGG16 Grad-CAM
Thresholded + Smoothed
Original | no_tumor
VGG16 Grad-CAM
Thresholded + Smoothed


ResNet50
Original | glioma_tumor
ResNet50 Grad-CAM
Thresholded + Smoothed
Original | meningioma_tumor
ResNet50 Grad-CAM
Thresholded + Smoothed
Original | pituitary_tumor
ResNet50 Grad-CAM
Thresholded + Smoothed
Original | no_tumor
ResNet50 Grad-CAM
Thresholded + Smoothed


EfficientNetB0
Original | glioma_tumor
EfficientNetB0 Grad-CAM
Thresholded + Smoothed
Original | meningioma_tumor
EfficientNetB0 Grad-CAM
Thresholded + Smoothed
Original | pituitary_tumor
EfficientNetB0 Grad-CAM
Thresholded + Smoothed
Original | no_tumor
EfficientNetB0 Grad-CAM
Thresholded + Smoothed


Vision Transformer (ViT)
Original | glioma_tumor
ViT Grad-CAM
Thresholded + Smoothed
Original | meningioma_tumor
ViT Grad-CAM
Thresholded + Smoothed
Original | pituitary_tumor
ViT Grad-CAM
Thresholded + Smoothed
Original | no_tumor
ViT Grad-CAM
Thresholded + Smoothed

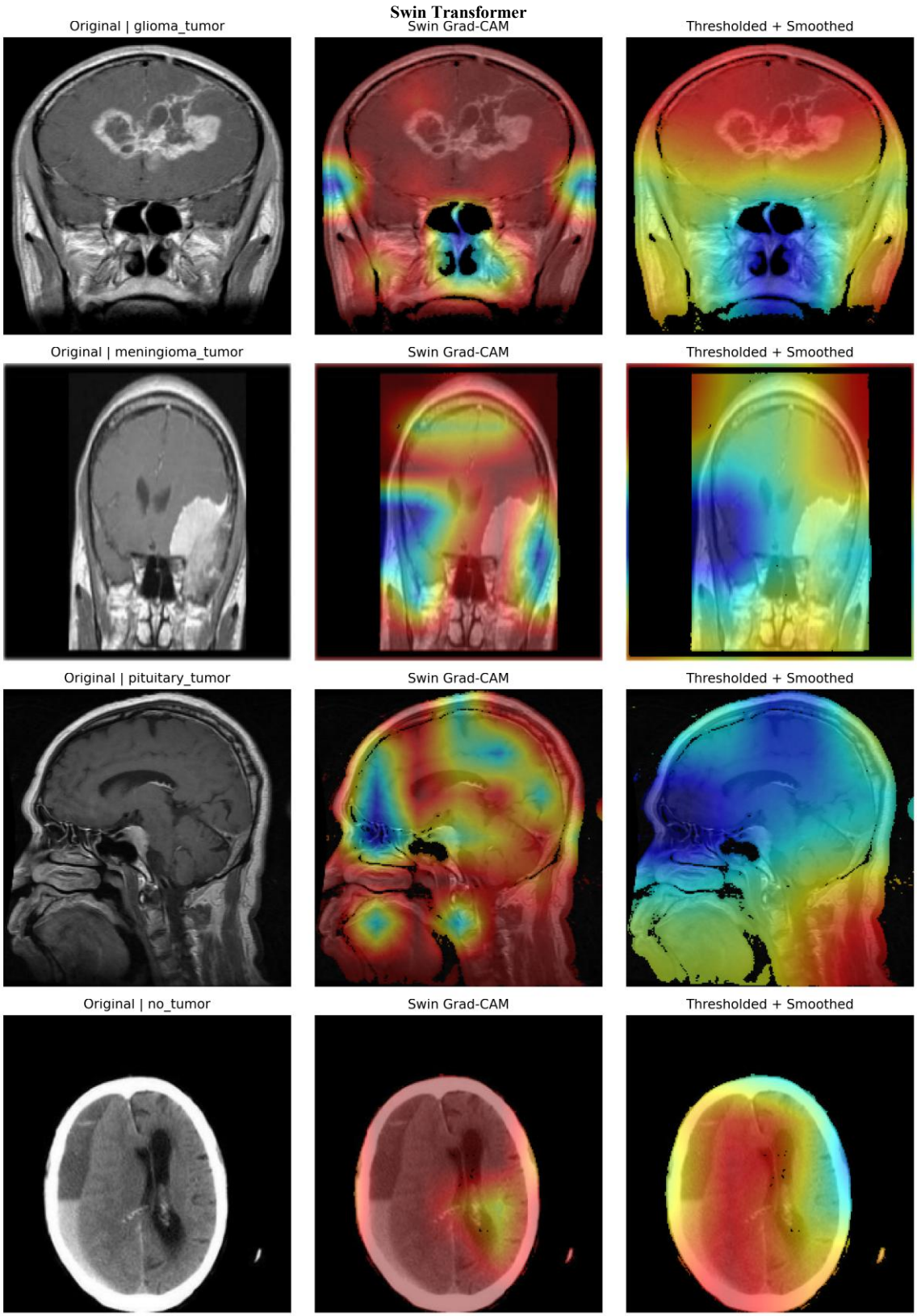


Fig. 8 Grad-CAM visualizations for CNN and transformer models across four classes (glioma, meningioma, pituitary tumor, and no-tumor). Each panel shows the original image, Grad-CAM overlay, and the post-processed (thresholded + smoothed) map. Interpretability patterns are evaluated qualitatively and with caution [11]

5. Discussion

The present study evaluated convolutional neural network and vision transformer models for four-class brain tumor MRI classification using a reliability- and explainability-oriented assessment framework. The results demonstrate that model performance in medical image classification should not be interpreted only through overall accuracy. Although EfficientNetB0 achieved the highest classification accuracy, macro-F1-score, and micro-average AUC, the broader evaluation indicates that class-wise behaviour, calibration, uncertainty, confusion patterns, and explanation quality provide additional information that is necessary for clinical decision-support applications. Therefore, the findings should be interpreted from both a predictive and clinical reliability perspective.

The discussion is organized around the major concerns raised in the evaluation. Section 5.1 discusses the relationship between classification performance and error structure. Section 5.2 interprets reliability, calibration, and uncertainty behaviour. Section 5.3 discusses explainability and the need for quantitative validation. Section 5.4 explains the importance of multi-modal and multi-sequence MRI. Section 5.5 discusses clinical workflow integration, robustness, and deployment. Section 5.6 highlights the role of human factors and radiologist-centered validation. Section 5.7 compares the present findings with recent state-of-the-art directions and clarifies the originality of the study.

5.1. Performance Versus Error Structure

The classification results indicate that the difference between the evaluated architectures is not limited to overall accuracy. As reported in Figure 1 and Table 1, EfficientNetB0 achieved the strongest overall performance among the five models, with an accuracy of 0.871, macro-precision of 0.861, macro-recall of 0.854, macro-F1-score of 0.857, and micro-average AUC of 0.941. Swin-Base obtained the second-best performance, followed by ViT-Base, ResNet50, and VGG16. This ranking suggests that both modern compound-scaled CNNs and hierarchical transformer models are more effective than older convolutional baselines for the present four-class brain MRI classification task.

EfficientNetB0 scores higher because to its compound scaling, which balances network depth, width, and input resolution. This allows the model to learn discriminative local texture, tumor boundary, and intensity-related features from 2D MRI slices without introducing excessive model complexity. Comparatively, VGG16 uses a relatively old sequential convolutional architecture, which can restrict its capacity to detect the intricate morphology of lesions. ResNet50 is better than VGG16 with residual connections, but it is still worse than EfficientNetB0, which suggests that deeper learning in features is not enough when tumor classes exhibit similar visual features.

The models based on transformers also delivered competitive performance. Swin-Base outperformed ViT-Base, which could be explained by its hierarchical shifted-window attention. In contrast to standard ViT, which learns global relations based on fixed patch tokens, Swin-Base learns local and contextual information by modeling the attention on windows at different levels. This can be helpful in brain MRI classification as the appearance of tumors can be different in terms of shape, texture, delimiting the boundaries and the surrounding tissues. The relatively small difference between EfficientNetB0 and Swin-Base also suggests that both strong local feature extraction and hierarchical attention are beneficial for tumor subtype discrimination.

Nevertheless, the per-class outcomes in Figure 2 and the confusion-matrix analysis in Figure 3 indicate that the classification issue is still a clinical challenge. Pituitary tumor and no-tumor cases were found to be more consistently identified, and glioma and meningioma created more difficulty by class. The significance of this pattern is that glioma and meningioma can exhibit overlapping intensity patterns, irregular morphology, and irregular boundaries in 2D axial MRI slices. Thus, the errors seen are not only due to the weaknesses of the model but also due to the limited diagnostic data that can be obtained in single-slice, single-sequence imaging.

The structure of error also demonstrates the fact that the aggregate accuracy is not enough to determine clinical usefulness. A model can be highly accurate in general, but still result in clinically significant confusion between tumor subtypes. In real-life neuro-oncology processes, this confusion can be applied to diagnostic prioritization, surgical planning, radiotherapy planning, or further imaging. Thus, the findings must be seen as an indication of better classification performance as well as an indication that more clinically complete inputs are necessary in future development.

Overall, Section 5.1 reinforces the interpretation of the original results, explaining why EfficientNetB0 and Swin-Base were more effective, why glioma-meningioma confusion remained, and why the next step should be taken to multi-sequence MRI, 3D volumetric modelling, external validation, and integration of clinical metadata. This interpretation is a direct reflection of the reviewer's demand to further analyze and explain how the proposed evaluation is better than the traditional accuracy-based comparisons.

5.2. Comparison with Recent Transformer Variants

Recent transformer-based models have reported accuracy values close to 99% on similar four-class brain tumor MRI classification tasks, showing a clear improvement over standard baseline models. For tumor size and appearance, the HMSA form uses multi-scale patch embedding and cross-attention. After temperature scaling, it had 98.7% accuracy and 0.023 Expected Calibration Error [9, 15]. This indicates

that multi-scale attention and post-hoc calibration can improve both discriminative performance and confidence reliability.

The FCM-SCA model combines feature calibration with selective cross-attention and has attained an accuracy of 98.9% and an F1-score of 99.23% [16]. These findings indicate that attention-based calibration modules have the potential to enhance separation of classes by fine-tuning the feature representation prior to ultimate classification. Likewise, fine-tuned vision transformer models like FTVT have demonstrated better performance in comparison to the traditional CNN baselines with an accuracy of 98.70% reported [17]. These results suggest that the performance of transformers can be significantly enhanced in the case of pre-trained attention models that are skillfully translated to brain MRI classification.

Pacal et al. conducted a systematic comparison of different transformer families and reported that Swin-Small achieved 99.37% accuracy while being more efficient than Swin-Large [18]. Their results also revealed that 0.1 label smoothing enhanced calibration without deteriorating classification performance [18]. This observation is relevant to the present study because label smoothing was also included in the training protocol to reduce overconfidence during classification.

The results of these recent studies show that standard ViT models can be strengthened through multi-scale embeddings, cross-attention, feature calibration, selective attention, and improved calibration strategies. However, these improvements often increase computational cost because they may require additional parameters, multi-branch attention blocks, or multiple input patch scales. For example, although Swin-Large may provide higher representational capacity, it can require substantially more parameters than smaller Swin variants, while HMSA extracts patches at three scales to improve feature diversity [15, 18].

In comparison, the present study does not claim to outperform these highly optimized transformer variants. Instead, it provides a controlled, reliability-oriented benchmark of representative CNN and transformer backbones under a common training and evaluation protocol. The results show that EfficientNetB0 achieved the best performance in the present benchmark, with an accuracy of 0.871, macro-F1-score of 0.857, AUC of 0.941, and ECE of 0.164, while Swin-Base achieved an accuracy of 0.865, AUC of 0.938, and ECE of 0.195. These baseline values are lower than the best reported transformer variants, but they are useful because they connect classification performance with calibration, uncertainty, and explainability analysis.

Therefore, the comparison with recent transformer variants highlights an important trade-off among accuracy, calibration, interpretability, and computational efficiency. Highly optimized transformer models may achieve superior accuracy, but their clinical value still depends on calibration, robustness to domain shift, uncertainty estimation, explainability validation, and deployment feasibility. The present benchmark should therefore be interpreted as a foundation for future extensions using multi-scale attention, feature calibration, selective cross-attention, and multi-sequence MRI inputs.

5.3. Reliability Considerations for Clinical Use

Calibration analysis shows that the reliability of predicted confidence differs across architectures. This is clinically important because an overconfident, incorrect prediction may be more harmful than an uncertain prediction. In medical decision-support systems, uncertainty can encourage expert review, whereas overconfidence may suppress referral to a radiologist or delay further diagnostic investigation. Therefore, the confidence score of a brain tumor MRI classifier should not be interpreted as clinically meaningful unless the model has been explicitly evaluated for calibration [8, 9].

A well-calibrated model can support safer triage policies. In such a workflow, low-confidence cases may be flagged or deferred for expert assessment, while high-confidence cases may be prioritized for faster review. This is particularly relevant in brain tumor classification because false-negative no-tumor predictions and incorrect tumor-subtype predictions may have different clinical consequences. Reliable confidence estimation, therefore, helps connect model output with practical clinical action rather than treating classification as a purely computational task [8, 9].

Post-hoc calibration techniques can also be used prior to deployment when the forecasted probabilities are not accurate enough. Probability scaling: Temperature scaling and isotonic regression are generally employed to enhance the reliability of predictions without necessarily altering the predicted class. These techniques come in handy when a model has high discrimination with poor calibration of confidence.

The high calibration in the recent studies of transformers also suggests that refinement of the calibration could enhance the clinical utility of model confidence scores [9, 15]. Another path to safer deployment is through uncertainty-conscious approaches. “Deep ensembles”, “Monte Carlo Dropout”, and “Bayesian neural networks” can estimate the level of belief or uncertainty of predictions. These techniques prove to be effective as they may assist in detecting unclear images, low-quality slices, scanner-shifted inputs, or out-of-distribution cases. It has also been demonstrated in previous research that uncertainty-sensitive methods are able to minimize false positives without compromising classification accuracy. These techniques can be combined with transformer variants to potentially generate more robust clinical models.

Reliability metrics also provide flexibility in selecting the operating point of the system. For example, if a hospital or radiology workflow requires the false-negative rate to remain below a maximum tolerated level, confidence thresholds can be adjusted accordingly. However, such threshold-based decision-making is meaningful only when probability outputs are well calibrated. In this sense, calibration analysis is not merely descriptive; it directly supports deployable strategies such as selective prediction, deferral, and triage prioritization [8].

Overall, calibration and uncertainty estimation are essential for translating brain tumor MRI classifiers into clinical decision-support workflows. A clinically useful model should not only predict the correct class but should also indicate when its prediction is uncertain or should not be trusted without expert review. Such selective prediction workflows, where low-confidence cases are deferred to radiologists, are increasingly recognized as an important requirement for safe clinical integration of artificial intelligence.

5.4. Explainability and Saliency-Map behavior

Grad-CAM comparisons provide qualitative evidence of the image regions used by the models during prediction. In the present study, the saliency maps indicate that modern backbones generally produced more lesion-aligned activation patterns, whereas classical baselines showed comparatively diffuse activations in some cases. Lesion-aligned saliency is consistent with the possibility that the model is learning pathology-relevant features, while diffuse activation may indicate reliance on broader anatomical context or non-specific image patterns. However, saliency maps should be interpreted cautiously because they are sensitive to preprocessing, selected layers, visualization settings, and implementation details. Thus, Grad-CAM must be viewed as an aid to model behaviour, not as conclusive evidence of proper clinical thinking [7, 11]. The multi-model Grad-CAM comparison also indicates that the consistency of explanations can be used as an informal measure of reliability. The more architectures indicate the same region of lesions as related to the same pattern of class, the more evidence that the decision is made on similar visual cues. Conversely, a high degree of diffuseness or non-uniformity of saliency across architectures can be a sign of the requirement for more detailed auditing, preprocessing verification, or supplementary explanation mechanisms. This is of particular concern in medical imaging since a visually plausible prediction can still be made on irrelevant background features. To further explainability than qualitative visualization, quantitative validation must be added, where it has expert annotations. Where tumor mask or radiologist-labeled areas are accessible, thresholded Grad-CAM maps may be compared to expert-defined lesion areas by Dice coefficient, Intersection over Union, localization agreement, or overlap ratio. These metrics can help determine whether the highlighted regions actually correspond to clinically meaningful tumor areas rather than unrelated anatomical structures. He directly addresses the limitation of relying only on visual inspection of saliency maps. Recent explainability literature also suggests that saliency-based methods should be used with other types of explainability, such as concept-based explanations, occlusion sensitivity, Integrated Gradients, Score-CAM, and counterfactual visualizations. These techniques may yield complementary data on whether the model relies on shape, texture, boundary features, enhancement patterns, or tissue context. A combination of these strategies with the current framework would give a more comprehensive view of model reasoning. In general, the Grad-CAM findings confirm the utility of explainability analysis in the context of comparing CNN and transformer models, but more stringent validation is needed. Future research ought to shift to more qualitative inspection of saliency to expert-validated and quantitative explainability. This would render the interpretation more clinically significant and would aid in establishing whether the high-performing models are focusing on tumor-relevant regions in reality.

5.5. Limitations and Future Work

This paper has some limitations. A single dataset of 7023 axial MRI slices supports the results. Although this dataset provides a useful basis for controlled benchmarking, external validation using multi-center and multi-vendor MRI data is required to confirm generalization. Variations in scanner vendor, magnetic field strength, acquisition protocol, slice thickness, image contrast, and institutional imaging practice may affect model performance. Therefore, future work should evaluate the proposed framework on independent datasets collected from different hospitals and scanner environments.

Second, only representative CNN and transformer backbones were evaluated in the present study. Recent multi-scale and calibration-aware transformer variants, including HMSA, FCM-SCA, and Swin-Small, have reported improved accuracy and calibration behaviour. Including these architectures within the same benchmarking framework would provide a more complete comparison between classical CNNs, standard transformers, hierarchical transformers, and calibration-aware attention models.

Third, Grad-CAM was used as a post-hoc explainability technique. Although it provides useful visual evidence of model attention, it may be sensitive to preprocessing, selected layers, model architecture, and visualization settings. Therefore, Grad-CAM may not capture all failure modes and should not be interpreted as definitive proof of clinical reasoning. Future work should include complementary explainability techniques such as concept-based explanations, counterfactual visualizations, Integrated Gradients, Score-CAM, and occlusion sensitivity, along with quantitative sanity checks [11]. Fourth, explicit uncertainty modelling and post-hoc calibration methods were not explored in full detail.

Although MC Dropout summaries were reported to estimate uncertainty, architecture-specific dropout placement or alternative Bayesian approximations may be required to obtain meaningful variability estimates, especially for transformer-based models. Future work should compare MC Dropout with "Bayesian neural networks", "deep ensembles", "evidential learning", "temperature scaling", and "isotonic regression". These approaches may provide additional reliability for borderline, ambiguous, low-quality, or out-of-distribution MRI cases [22].

Fifth, the use of 3D context and multi-sequence MRI is a natural extension for reducing glioma–meningioma confusion. This study uses 2D single-sequence MRI slices, which may not capture volumetric tumor continuity, inter-slice information, edema distribution, enhanced characteristics, or multi-sequence radiological evidence. Future models should incorporate T1, contrast-enhanced T1, T2, FLAIR, and relevant clinical metadata wherever available. This direction is consistent with recent ViT-BT and FTVT studies that exploit multi-modal or multi-sequence MRI information [17, 19].

Sixth, the visualizations of Grad-CAM reported are mostly qualitative. Expansion into expert-annotated tumor masks in the future will enable quantitative assessment of saliency maps based on quantitative localization measures like Dice coefficient, Intersection over Union, and overlap ratio, as well as localization agreement. This would render the explainability analysis more objective and clinically meaningful. Lastly, calibration behaviour can vary when using a model with data from new scanners, institutions, or patient groups. Thus, it should be deployed with continuous monitoring, recalibration periodically, and strength checks in the case of a distribution shift. Radiologist-centered assessment should also be considered in the future to understand whether model predictions, confidence scores, uncertainty indicators, and explanation maps can be helpful in actual reporting processes.

In general, future studies must transition to externally validated, uncertain, quantitatively explainable, multi-sequence, and clinically integrated brain tumor MRI analysis, rather than just single-dataset classification benchmarking. This trend would enhance classification performance, reliability, interpretability, and deployment readiness.

6. Conclusion

This study presented a reliability- and explainability-oriented comparison of CNN and vision transformer models for brain tumor MRI classification. In addition to conventional classification performance, the evaluation considered calibration through reliability diagrams and Expected Calibration Error, as well as interpretability through post-processed Grad-CAM visualizations. All models were evaluated under a common training protocol, allowing a controlled comparison of discrimination, confidence, reliability, and explanation behaviour. The results indicate that modern deep learning backbones provide advantages not only in classification performance but also in clinically relevant reliability characteristics. The overall performance of EfficientNetB0 was the best in the current benchmark, and Swin-Base demonstrated competitive behaviour in terms of transformer-based models. Nevertheless, it is also revealed that the enhancement of one of the evaluation dimensions does not necessarily lead to the enhancement of another. Before being used for clinical decision-making, even a good model must be thoroughly evaluated for calibration, uncertainty, and saliency-map behavior. The results were also contextualized in terms of the recent multi-scale and calibration-aware transformer models. More sophisticated forms of transformers have been reported to be more accurate and better calibrated, though they can also be more computationally expensive and architecturally complex. Therefore, the results highlight an important trade-off among accuracy, calibration, interpretability, and deployment feasibility. The current benchmark offers a consistent starting point to future research that could combine calibration-conscious aims, uncertainty quantification, multi-scale focus, and transformer designs that are specifically tailored to medical imaging.

Overall, the study supports the need to evaluate biomedical AI systems through a multidimensional framework rather than through accuracy alone. A clinically useful brain tumor MRI classifier should be discriminative, well calibrated, interpretable, uncertainty-aware, and robust across clinical settings. Such a holistic evaluation is essential for narrowing the gap between proof-of-concept models and decision-support tools that can be trusted in real clinical workflows. Future work should therefore focus on uncertainty-aware training, quantitative explainability validation, external multi-center testing, multi-sequence MRI integration, and radiologist-centered evaluation.

References

[1] Ujjwal Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification," *arXiv Preprint*, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Maria Correia de Verdier et al., "The 2024 Brain Tumor Segmentation Challenge: Glioma Segmentation on Post-Treatment MRI," *arXiv*, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ashish Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar] [Publisher Link]

[4] Ze Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992-10002, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alexey Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv*, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Omid Nejati Manzari et al., "MedViT: A Robust Vision Transformer for Generalized Medical Image Classification," *Computers in Biology and Medicine*, vol. 157, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Bas H.M. van der Velden et al., "Explainable Artificial Intelligence in Deep Learning-based Medical Image Analysis," *Medical Image Analysis*, vol. 79, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ke Zou et al., "A Review of Uncertainty Estimation and its Application in Medical Imaging," *Meta-Radiology*, vol. 1, no. 1, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Chuan Guo et al., "On Calibration of Modern Neural Networks," *Proceedings of the 34th International Conference on Machine Learning, PMLR*, vol. 70, pp. 1321-1330, 2017. [Google Scholar] [Publisher Link]

[10] Amna Iqbal, Muhammad Arfan Jaffar, and Rashid Jahangir et al., "Enhancing Brain Tumour Multi-Classification using Efficient-Net B0-based Intelligent Diagnosis for Internet of Medical Things (IoMT) Applications," *Information*, vol. 15, no. 8, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Katarzyna Borys et al., "Explainable AI in Medical Imaging: An Overview for Clinical Practitioners – Beyond Saliency-based XAI Approaches," *European Journal of Radiology*, vol. 162, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jieneng Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ali Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574-584, 2022. [Google Scholar] [Publisher Link]

[14] Shakif Ahmed et al., "Advancing Brain Tumor MRI Classification using SwRD: A Parallel Swin Transformer-ResNet Approach," *Virtual Reality and Intelligent Hardware*, vol. 7, no. 5, pp. 501-522, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] C. Sankari, V. Jamuna, and A.R. Kavitha et al., "Hierarchical Multi-Scale Vision Transformer Model for Accurate Detection and Classification of Brain Tumors in MRI-based Medical Imaging," *Scientific Reports*, vol. 15, no. 1, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mohammad Ali Labbaf Khaniki et al., "Vision Transformer with Feature Calibration and Selective Cross-Attention for Brain Tumor Classification," *Iran Journal of Computer Science*, vol. 8, no. 2, pp. 335-347, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] C. Kishor Kumar Reddy et al., "A Fine-Tuned Vision Transformer based Enhanced Multi-Class Brain Tumor Classification using MRI Scan Imagery," *Frontiers in Oncology*, vol. 14, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Yigitcan Cakmak, and Ishak Pacal, "Comparative Analysis of Transformer Architectures for Brain Tumor Classification," *Exploration of Medicine*, vol. 6, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Khawla Hussein Ali, "ViT-BT: Improving MRI Brain Tumor Classification using Vision Transformer with Transfer Learning," *International Journal of Soft Computing and Engineering*, vol. 14, no. 4, pp. 16-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ramprasaath R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017. [Google Scholar] [Publisher Link]

[21] Haomin Chen et al., "Explainable Medical Imaging AI needs Human-Centered Design: Guidelines and Evidence from a Systematic Review," *npj Digital Medicine*, vol. 5, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ling Huang et al., "A Review of Uncertainty Quantification in Medical Image Analysis: Probabilistic and Non-Probabilistic Methods," *Medical Image Analysis*, vol. 97, 2024. [CrossRef] [Google Scholar] [Publisher Link]