Journal of Associated Medical Sciences 2026; 59(2): 44-62

JAMS
Journal of Associated Medical Sciences

Scopus Indexed & Thai-Journal Citation Index Centre (TCI)

Journal of Associated Medical Sciences

Journal homepage: https://www.tci-thaijo.org/index.php/bulletinAMS/index

Comparative evaluation and interpretability analysis of modern CNN architectures for brain tumor MRI classification

Nitipon Pongphaw* and Prommin Buaphan

Department of Electrical and Computer Engineering, Faculty of Science and Engineering, Kasetsart University Chalermphrakiat Sakon Nakhon Province Campus, Sakon Nakhon, Thailand.

ARTICLE INFO

Article history:
Received 3 November 2025
Accepted as revised 14 January 2026
Available online 23 January 2026

Keywords:
Brain tumor classification, magnetic resonance imaging, comparative study, transfer learning, grad-CAM.

ABSTRACT

Background: Accurate and interpretable brain tumor classification from MRI images remains a key challenge in medical image analysis, particularly when using publicly available datasets of moderate size.

Objective: This study investigates the performance of a ConvNeXt-Tiny based framework for four-class brain tumor classification glioma, meningioma, pituitary tumor, and no tumor and compares it with established convolutional architectures.

Materials and methods: Using transfer learning and identical experimental settings, ConvNeXt-Tiny was evaluated against DenseNet169, Xception, MobileNetV3-Large, CNN+DenseNet169, and ResNet50. Standard evaluation metrics (accuracy, precision, recall, and F1-score) were used, and Grad-CAM was applied to visualize model attention for interpretability. Generalization was further assessed using an independent dataset.

Results: ConvNeXt-Tiny achieved high overall performance (accuracy = 0.9924, F1-score = 0.9918), comparable to DenseNet169 and Xception but with lower computational cost. The model maintained stable learning behavior, minimal overfitting, and consistent accuracy on unseen data. Grad-CAM visualizations confirmed that the network focused on clinically relevant tumor regions, improving transparency and reliability of predictions.

Conclusion: ConvNeXt-Tiny provides a strong and efficient baseline for interpretable brain tumor classification, balancing accuracy and computational efficiency. While the results are promising, differences among recent architectures were modest, and clinical validation using multi-center MRI datasets is necessary to confirm broader applicability.

Introduction

The classification of brain tumors using deep learning techniques has become a transformative approach in medical imaging, particularly in magnetic resonance imaging (MRI). This progress is largely driven by the capability of deep learning models, especially Convolutional Neural Networks (CNNs), to automatically extract features from images and improve diagnostic accuracy.[1,2] Recent studies have demonstrated that integrating advanced architectures, such as residual networks and attention mechanisms, can enhance performance in differentiating various types of brain tumors, including gliomas and meningiomas.[3,4] Transfer learning has been widely

* **Corresponding contributor.**
Author's Address: *Department of Electrical and Computer Engineering, Faculty of Science and Engineering, Kasetsart University Chalermphrakiat Sakon Nakhon Province Campus, Sakon Nakhon, Thailand.*
E-mail address: nitipon.p@ku.th
doi: 10.12982/JAMS.2026.036
E-ISSN: 2539-6056

adopted to further improve classification accuracy, with models pre-trained on large datasets being fine-tuned on specific radiological datasets to better capture the nuances of brain tumor images.[1,5,6] For example, the use of ResNet50 and other architectures has shown significant improvements in metrics such as accuracy and F1-score.[2,7] Moreover, deep learning approaches have been reported to outperform traditional diagnostic methods, which often rely heavily on subjective interpretation.[8,9] Numerous studies have also explored model optimization strategies. Hybrid models combining CNNs with other machine learning classifiers, such as SVMs and decision trees, have shown improvements in both performance and generalizability.[10,11] Architectural innovations, including attention blocks and multi-tiered networks, aim to reduce overfitting while enhancing model interpretability.[3,12] Additionally, metaheuristic optimization techniques have been applied for feature selection to strengthen classification robustness.[13] Context-aware deep learning frameworks have further enriched the capability to segment and classify brain tumors while also predicting patient survival rates.[14] Multi-faceted strategies of this nature have demonstrated significant improvements in clinical outcomes, providing healthcare practitioners with effective tools for brain tumor diagnosis and management.[15] Ongoing research continues to advance deep learning applications in brain tumor classification, highlighting the potential for future clinical adoption. The ability to achieve high accuracy while handling the complex and heterogeneous nature of brain tumors underscores the critical role of deep learning in enhancing diagnostic workflows and patient care. Explainable Artificial Intelligence (XAI) techniques enhance model transparency and interpretability, allowing clinicians to understand the decision-making process behind deep learning predictions. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) and Gradient-weighted Class Activation Mapping (Grad-CAM) have been employed to explain how CNNs classify tumor types and detect tumor presence. For instance, Ullah *et al*. proposed DeepEBTDNet, which effectively utilized LIME to interpret MRI-based predictions, achieving a validation accuracy of 98.96%.[16] Similarly, Esmaeili *et al*. reported that incorporating interpretable methods could improve the accuracy of trained models.[17] Grad-CAM has been applied to visualize image regions most influential to the model's decisions. Goyal and Sharma used Grad-CAM to highlight features driving brain tumor detection, enhancing trust in AI-assisted diagnostics.[18] Tan *et al*. demonstrated the practical application of these visualization techniques in clinical workflows, improving diagnostic reliability and addressing concerns about the "black-box" nature of deep learning models, though reference[19] does not directly support Grad-CAM application in this context. Beyond interpretability, these techniques also enhance model robustness and reliability by showing which features are considered important, enabling practitioners to validate them against established medical knowledge.[20] Furthermore, attention mechanisms incorporated into deep learning frameworks allow models to focus on clinically relevant regions in MRI scans, which is critical for tumor segmentation and classification. Tonni *et al*. introduced a hybrid transfer learning framework incorporating attention-based features to improve both interpretability and performance.[19] Despite significant advances in deep learning-based brain tumor classification, several research gaps remain that limit model robustness and clinical applicability. One major gap is the reliance on large, well-annotated datasets for effective model training. Many studies rely on publicly available datasets that vary in size, balance, and acquisition conditions, which may lead to overfitting and reduced generalization when models are applied to unseen or heterogeneous MRI scans.[2] This highlights the importance of evaluating deep learning frameworks under realistic data availability and heterogeneous imaging conditions, while avoiding over-reliance on excessively large or curated clinical datasets. Another critical gap is the lack of interpretability of deep learning models in clinical practice. Although CNN-based architectures achieve high accuracy, their "black-box" nature makes it difficult for clinicians to understand the rationale behind predictions, reducing trust and hindering adoption in real-world healthcare settings.[20]

Materials and methods

This study proposes a deep learning framework for brain tumor classification from MRI images. The overall workflow is illustrated in Figure 1.

The methodology consists of five main components:

1. Proposed approach: This study presents a fair experimental comparison between ConvNeXt-Tiny as the baseline model and other deep learning architectures for four-class brain tumor classification. Beyond multi-class prediction, the proposed approach provides additional insights through subclass-level tumor analysis. Furthermore, binary classification between tumor and no-tumor cases is investigated to offer a clearer understanding of the model's diagnostic behavior.

2. Model configuration: The model performance was also evaluated in comparison with several CNN architectures, including ResNet50, CNN+DenseNet169, DenseNet169, MobileNetV3-Large, and Xception, using transfer learning.

3. Data preprocessing: The images were organized into training, validation, and test sets. All images were resized to 224×224 pixels prior to model training.

4. Evaluation: Performance of all models was assessed using standard metrics, including accuracy, precision, recall, and F1-score. Confusion matrix was used to visualize class-wise performance. In addition, training efficiency was evaluated by recording the

total training time and the average training time per epoch. These measurements provide insight into the computational cost of each model, allowing for a holistic comparison of predictive performance and training efficiency.

5. Grad-CAM: To provide insight into model decisions, Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to the trained models. Grad-CAM generates heatmaps highlighting the regions of input images that contribute most strongly to the predicted class. This allows visual assessment of whether the model focuses on relevant tumor regions, supporting interpretability and trustworthiness of the predictions.

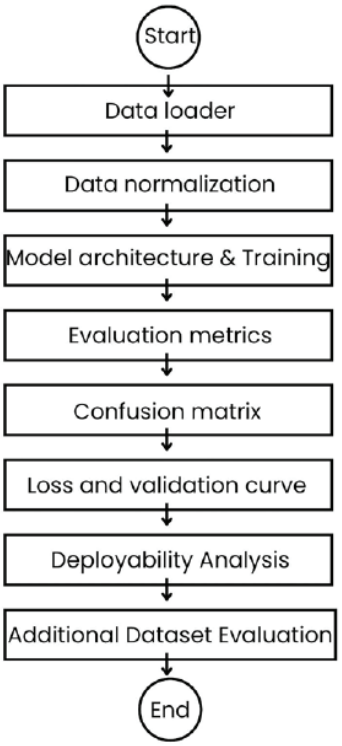

Figure 1. *Flowchart of methodology.*

Contributions

To address these challenges, this study makes two primary contributions. First, a fair and systematic comparison is conducted using ConvNeXt-Tiny as the baseline model to evaluate the performance of other deep learning architectures under identical training and evaluation conditions, thereby ensuring an objective assessment under identical experimental settings using a publicly available benchmark dataset. Second, the framework incorporates interpretability-enhancing mechanisms, including Grad-CAM analysis, to provide clear and clinically meaningful explanations for each classification decision. This integrated approach enhances diagnostic reliability while fostering clinical confidence and practical applicability in real-world brain tumor analysis.

Model architecture

ConvNeXt-Tiny was selected as the baseline model due to its modern convolutional architecture that bridges traditional CNNs with transformer-inspired design principles. By incorporating updated architectural choices, including larger kernel sizes, depthwise separable convolutions, and simplified normalization and activation strategies, ConvNeXt-Tiny enhances representational capacity while maintaining computational efficiency. The model was chosen a priori to balance expressive power and parameter efficiency, thereby reducing the risk of overfitting and remaining compatible with gradient-based interpretability techniques such as Grad-CAM. Larger ConvNeXt variants and transformer-based architectures were not prioritized because of their higher computational requirements and stronger dependence on large-scale training data.

Compared with lightweight models such as MobileNetV3-Large and deeper architectures including DenseNet169 and ResNet50, ConvNeXt-Tiny offers a well-balanced trade-off among model complexity, predictive performance, and generalization capability, making it a robust and reliable baseline for fair comparison. To ensure a comprehensive evaluation, six deep learning models were selected for comparison: CNN+DenseNet169, Xception, DenseNet169, ConvNeXt-Tiny, MobileNetV3-Large, and ResNet50. All models were trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and CrossEntropyLoss as the objective function on a four-class dataset. Detailed hyperparameter settings are summarized in Table 1, and schematic architectures of all models are illustrated in Figure 2.

Table 1. Parameter settings of all models.

Model	Parameters
CNN+DenseNet169	CNN(Conv2d(64) → BatchNorm → ReLU → MaxPool → Conv2d(128) → BatchNorm → ReLU → MaxPool) → Adapter (Conv2d 128→3) → Upsample (224×224) → DenseNet169(pretrained) → Dense(num_classes)
Xception	Xception(pretrained) → Dense(num_classes)
DenseNet169	DenseNet169(pretrained) → Dense(num_classes)
ConvNeXt-Tiny	ConvNeXt-Tiny(pretrained) → Dense(num_classes)
MobileNetV3-Large	MobileNetV3-Large(pretrained) → Dense(num_classes)
ResNet50	ResNet50(pretrained) → Dense(num_classes)

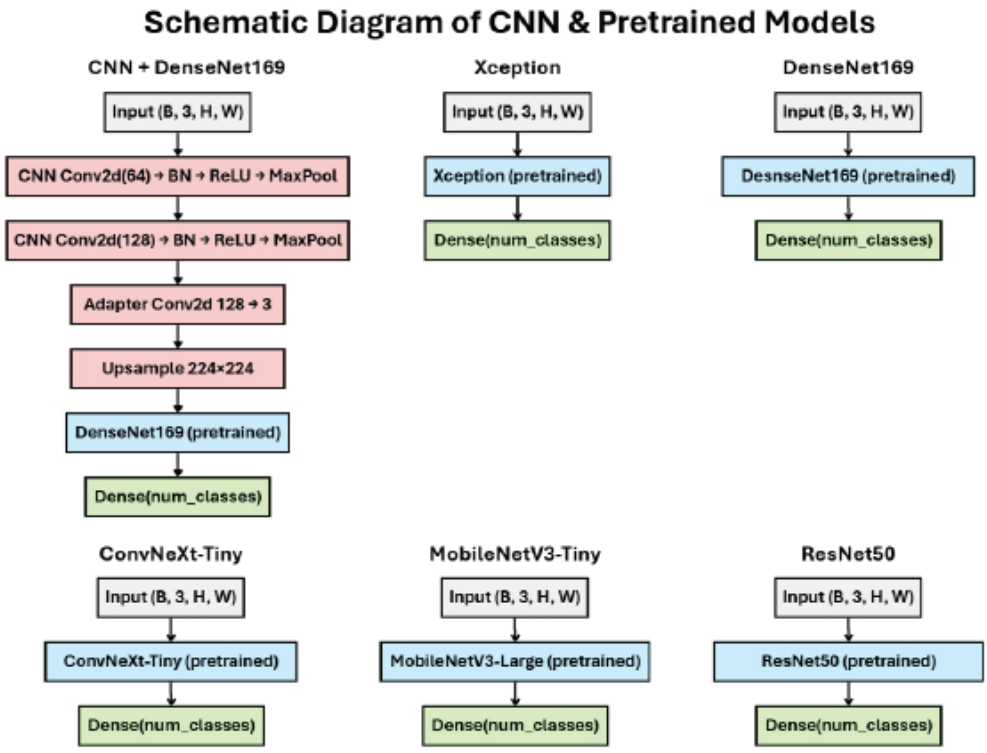

Figure 2. *Schematic diagrams of all models.*

Dataset

In this study, Dataset 1 was adopted from,[21] which is distributed under the CC0 license, and was used as the primary dataset for model training and initial evaluation. This dataset consists of 5,139 training images, 573 validation images, and 1,311 test images. For external validation, Dataset 2 was employed, namely the *Labeled MRI Brain Tumor Dataset Computer Vision Model* curated by Ali Rostami,[22] which is publicly available under the CC BY 4.0 license and contains a total of 2,443 MRI images. Dataset 2 was exclusively used to assess the generalization capability of the proposed model on unseen data. Detailed descriptions of both datasets, including class composition and source attribution, are provided in the Supplementary File. Figure 3 illustrates representative samples from each class in Dataset 1, while Figure 4 presents example images from Dataset 2.

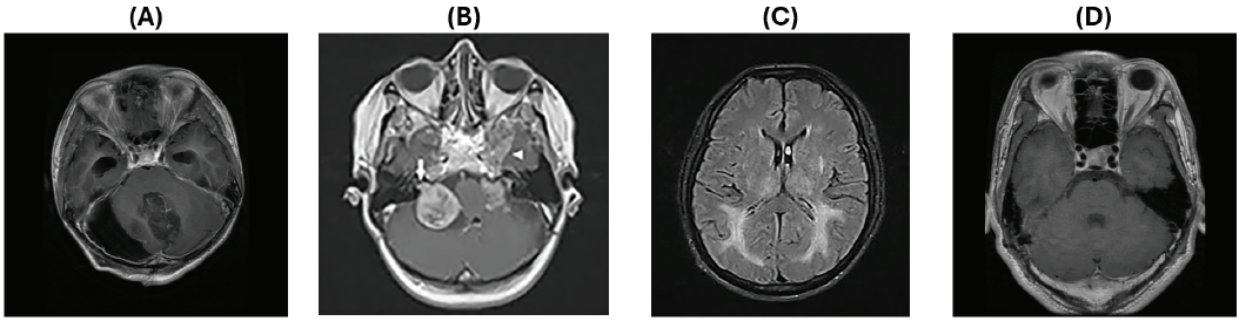

Figure 3. *Sample images from dataset 1. A: glioma, B: meningioma, C: no tumor, D: pituitary tumor.*

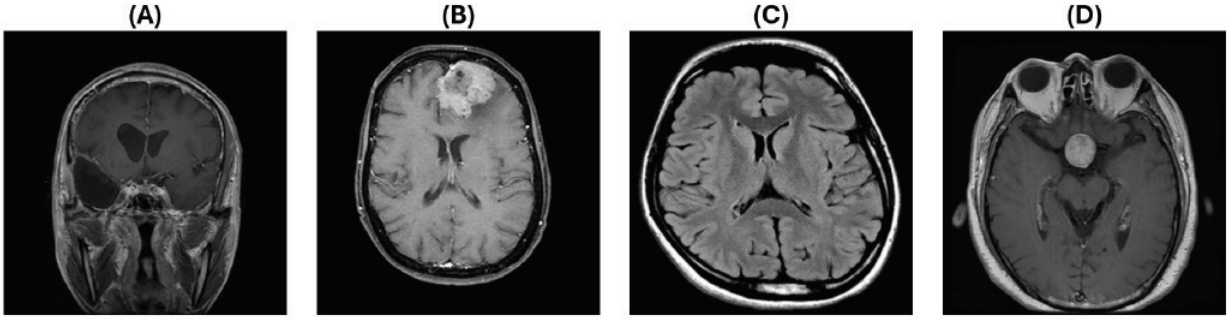

Figure 4. *Sample images from alternative dataset. A: glioma, B: meningioma, C: no tumor, D: pituitary tumor.*

Data preprocessing

In this study, all MRI images from both datasets were preprocessed before being used in the deep learning models. Dataset 1 was divided into training and validation sets using a 90:10 stratified split to preserve class balance, while all images in Dataset 2 were used exclusively for testing. All images were resized to 224×224 pixels, the standard input size for most convolutional neural networks (CNNs), to ensure consistency across architectures, and then converted into PyTorch tensors. The data were loaded using PyTorch's *ImageFolder* and *DataLoader* with a batch size of 32. Training data were shuffled at each epoch, whereas validation and test data were loaded sequentially. The final dataset included 5,139 training images, 573 validation images, and 1,311 test images from Dataset 1, together with all images from Dataset 2 used for external testing.

Evaluation metrics

To assess the performance of the proposed models, the following standard evaluation metrics were used: accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified images among all images. Precision indicates the proportion of correctly predicted positive samples among all samples predicted as positive. Recall represents the proportion of correctly predicted positive samples among all actual positive samples. F1-score is the harmonic means of precision and recall, providing a balance between the two. All metrics were calculated using the macro-average approach, which treats each class equally by computing the metric independently for each class and then averaging the results, regardless of class imbalance.

Computational resources

All experiments were executed on Google Colab Pro using an NVIDIA H100 GPU. Model training and inference were performed using CUDA acceleration, as indicated by the device configuration (using device: cuda).

Results

In this study, we hypothesized that the ConvNeXt-Tiny model could serve as an effective baseline for comparative evaluation of convolutional architectures in multi-class brain tumor classification. All models were trained and evaluated under identical experimental settings to ensure fair comparison.

Evaluation metrics

The performance of all evaluated models is summarized in Table 2 Overall, the results show that several architectures achieved comparably high performance under identical experimental settings. DenseNet169 yielded the highest accuracy (0.9969), followed closely by MobileNetV3-Large (0.9947) and Xception (0.9939). ConvNeXt-Tiny, used as the main baseline in this study, also demonstrated competitive results (accuracy=0.9924, precision=0.9920, recall=0.9917, F1-score=0.9918), indicating its reliability and efficiency for this task. CNN+DenseNet169 performed slightly lower but remained strong overall, while ResNet50 showed comparatively lower metrics across all measures. These findings suggest that multiple convolutional architectures can achieve similar performance levels in multi-class brain tumor classification, emphasizing the value of comparative evaluation rather than reliance on a single model.

Table 2. Performance of all models.

Model	Accuracy	Precision	Recall	F1-score
CNN+Densenet169	0.9916	0.9914	0.9909	0.9912
Xception	0.9939	0.9936	0.9934	0.9934
Densenet169	0.9969	0.9969	0.9967	0.9968
ConvNeXt-Tiny	0.9924	0.9920	0.9917	0.9918
MobileNetV3-Large	0.9947	0.9945	0.9942	0.9943
ResNet50	0.9680	0.9672	0.9661	0.9664

Loss and validation curve

Figure 5 illustrates the training and validation loss and accuracy curves of the baseline ConvNeXt-Tiny model. Both curves demonstrate a rapid stabilization during the early training phase, with the loss decreasing sharply and reaching a steady level around the fifth epoch, while the accuracy curves gradually converged and remained closely aligned throughout the remaining epochs. This behavior indicates a stable training process with no evident overfitting or divergence between training and validation performance. Compared with prior studies employing ConvNeXt-based or other deep CNN architectures for image classification, which often report longer training schedules ranging from 30 to over 100 epochs depending on the dataset and imaging modality, the convergence observed within 20 epochs in this study suggests relatively fast learning dynamics. This behavior is primarily attributed to the use of transfer learning, which enables the ConvNeXt-Tiny model to exploit pretrained feature representations and reduces the need for extensive parameter updates. In addition, the relatively consistent structural characteristics of brain MRI images may facilitate efficient feature adaptation compared with natural image datasets.

Training and validation loss and accuracy curves for baseline model are presented in Figure 5.

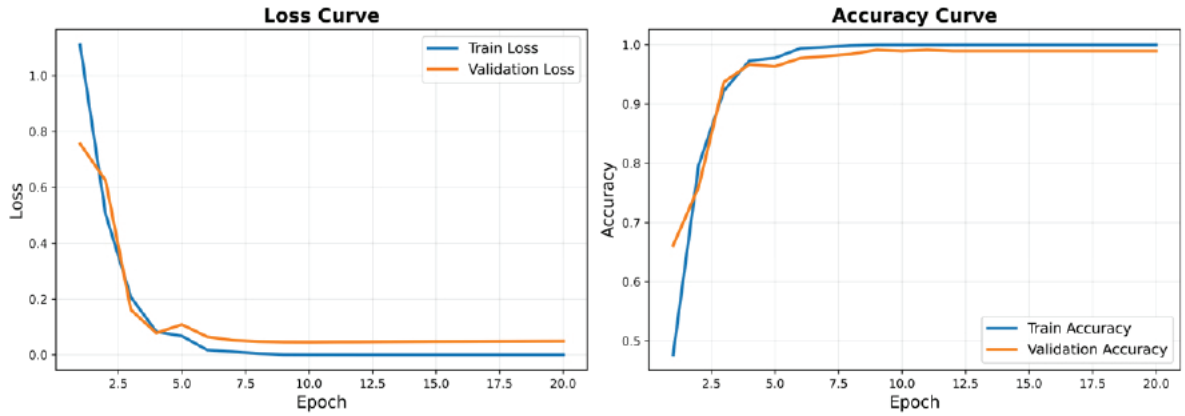

Figure 5. *Training and validation performance curves of the ConvNeXt-Tiny.*

Confusion matrix

To gain further insight into the classification performance of each model, a confusion matrix was constructed to illustrate class-wise predictions for the four brain tumor categories: glioma, meningioma, no tumor, and pituitary tumor.

Table 3 shows the confusion matrix of the baseline ConvNeXt-Tiny model on the test set. The model correctly classified most samples across all four categories, with only a few misclassifications observed between *glioma* and *meningioma*, and between *meningioma* and *pituitary*. The "no tumor" class was predicted with perfect consistency, indicating clear separability from tumor classes under the current dataset. These results suggest that the baseline model maintained balanced recognition among categories, with minor overlaps primarily occurring between histologically related tumor types.

Table 3. Confusion matrix of the ConvNeXt-Tiny model.

	Predicted glioma	Predicted meningioma	Predicted notumor	Predicted pituitary
Actual glioma	295	4	0	1
Actual meningioma	2	303	1	0
Actual no tumor	0	0	405	0
Actual pituitary	0	2	0	298

Learning curve

The learning curve (Figure 6) analysis indicates that both validation and test accuracy increase consistently with the number of training samples, demonstrating that the model benefits substantially from additional data. Early in training, with fewer than 200 samples, accuracy remains relatively low (~0.63-0.88), but a marked improvement is observed as the dataset grows to 500-1000 samples, reaching ~0.95. Beyond 1000 samples, gains in accuracy begin to plateau, indicating diminishing returns, with the model approaching its maximum performance at 5163 samples (~0.99). The close alignment between validation and test accuracy across all sample sizes suggests minimal overfitting and good generalization.

Stratified sampling was used to ensure each class contributed equally to the training subsets, preventing bias and supporting balanced learning across all classes.

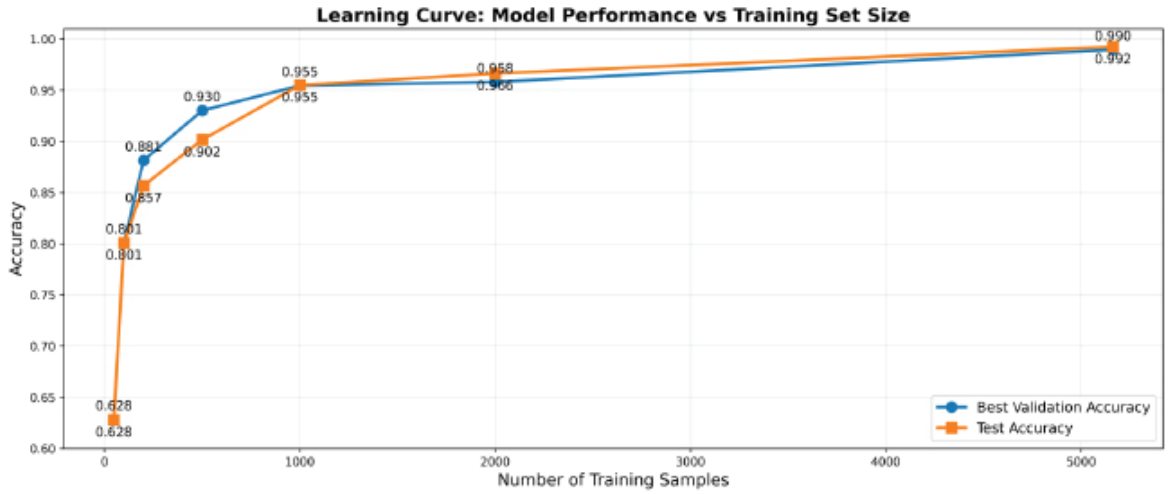

Figure 6. *Effect of training sample size on ConvNeXt-Tiny accuracy*

Statistical test

Null hypothesis (H_0)

There is no difference in the proportion of correctly classified samples between ConvNeXt-Tiny and the compared model.

Alternative hypothesis (H_1)

There is a significant difference in the proportion of correctly classified samples between ConvNeXt-Tiny and the compared model.

Table 4 summarized of the results of McNemar's tests comparing ConvNeXt-Tiny with other models indicate that for most comparisons, there is no statistically significant difference in performance. Specifically, comparisons with CNN+DenseNet169 (χ^2=0.0, p=1.0), Xception (χ^2=0.083, p=0.773), DenseNet169 (χ^2=3.125, p=0.077), and MobileNetV3 (χ^2=0.444, p=0.505) all yielded p>0.05, indicating that their performance differences with ConvNeXt-Tiny are not statistically significant. In contrast, the comparison with ResNet50 (χ^2=22.881, p<0.001) shows a highly significant difference, confirming that ConvNeXt-Tiny outperforms ResNet50. Overall, these results suggest that ConvNeXt-Tiny performs comparably to most modern architectures while significantly surpassing the older ResNet50 model.

Table 4. McNemar's test comparison of ConvNeXt-Tiny with other baseline models.

Baseline	Compared Model	Chi-square	*p* value	Significant (p<0.05)
ConvNeXt-Tiny	CNN+DenseNet169	0.0	1.0	No
ConvNeXt-Tiny	Xception	0.0833	0.7728	No
ConvNeXt-Tiny	DenseNet169	3.1250	0.0771	No
ConvNeXt-Tiny	MobileNetV3	0.4444	0.5050	No
ConvNeXt-Tiny	ResNet50	22.8810	1.7235×10^{-6}	Yes

The 95% confidence interval (CI)

According to Table 5, the six models show clear differences in both accuracy and Macro-F1. DenseNet169 achieved the highest performance (Accuracy 0.997±0.0053, Macro-F1 0.997±0.0054), indicating consistent and accurate classification across all classes. MobileNetV3 and Xception offer comparable results with low variability, suitable for lightweight or fast-inference applications. CNN+DenseNet169 and ConvNeXt-Tiny also performed well, slightly below DenseNet169. ResNet50 had the lowest accuracy and higher variability, suggesting a need for hyperparameter tuning or additional data augmentation. Overall, modern architectures provide superior accuracy and stability, with DenseNet169 representing the optimal choice for maximal classification performance.

Table 5. Model accuracy and Macro-F1 scores (95% CI).

Model	Accuracy (95% CI)	Macro-F1 (95% CI)
CNN + DenseNet169	0.992±0.0059	0.991±0.0055
Xception	0.994±0.0044	0.993±0.0051
DenseNet169	0.997±0.0053	0.997±0.0054
ConvNeXt-Tiny	0.992±0.0049	0.992±0.0051
MobileNetV3	0.995±0.0044	0.994±0.0045
ResNet50	0.968±0.0091	0.966±0.0090

Computational metrics

Table 6 summarizes the GPU memory usage and training time per epoch for all evaluated models. Among the models, MobileNetV3-Large required the least GPU memory (1,423.90 MB) and had the fastest training time per epoch (25.72 s), reflecting its lightweight architecture. ConvNeXt-Tiny, while achieving high overall performance, used moderate GPU memory (3,721.97 MB) and required 38.00 s per epoch for training. CNN+DenseNet169 consumed the most GPU memory (6,770.84 MB) with a training time of 40.10 s per epoch. Xception and DenseNet169 had moderate memory usage (4,419.10 MB and 5,037.70 MB, respectively) with training times of 30.19 s and 38.36 s per epoch. ResNet50 required 2,959.80 MB of GPU memory and 26.91 s per epoch. These results indicate that lightweight models like MobileNetV3-Large are highly efficient in terms of resource usage, while ConvNeXt-Tiny provides a good balance between classification performance and computational cost.

Table 6. Training time and memory usage of all models.

Model	GPU memory usage (MB)/epoch	Training time (s)/epoch
CNN + DenseNet169	6770.84	40.10
Xception	4419.10	30.19
DenseNet169	5037.70	38.36
ConvNeXt-Tiny	3721.97	38.00
MobileNetV3-Large	1423.90	25.72
ResNet50	2959.80	26.91

Subclass tumor

Table 7 and Table 8 demonstrate that ConvNeXt-Tiny achieved high and consistent metrics (accuracy 0.9923, precision 0.9940, recall 0.9923, F1-score 0.9931), indicating reliable identification of all subclasses. Specifically, the model correctly predicted glioma 298/300 (99.3%), meningioma 303/306 (99.0%), and pituitary 298/300 (99.3%), with only seven misclassifications among 906 test samples. The close alignment of precision, recall, and F1-score suggests balanced performance without bias toward any subclass, supporting its use as a baseline for further subclass classification studies.

Table 7. ConvNeXt-Tiny subclass prediction performance.

Model	Accuracy	Precision	Recall	F1-score
ConvNeXt-Tiny	0.9923	0.9940	0.9923	0.9931

Table 8. Confusion matrix (test set without no tumor).

	Predicted glioma	Predicted meningioma	Predicted pituitary
Actual glioma	298	2	0
Actual meningioma	2	303	1
Actual pituitary	0	2	298

No tumor vs tumor

Table 9 and Table 10 demonstrates that ConvNeXt-Tiny achieved very high and balanced performance (accuracy 0.9985, precision 0.9982, recall 0.9982, F1-score 0.9982) on the test set, which included 906 tumor samples and 405 no-tumor samples. The model correctly identified 905/906 tumor cases and 404/405 no-tumor cases, with minimal misclassifications. The close alignment of precision, recall, and F1-score indicates minimal bias toward either class, confirming its suitability as a reliable baseline model for binary classification with exceptional diagnostic performance.

Table 9. ConvNeXt-Tiny no tumor vs tumor prediction performance.

Model	Accuracy	Precision	Recall	F1-score
ConvNeXt-Tiny	0.9985	0.9982	0.9982	0.9982

Table 10. Confusion matrix (test set no tumor vs tumor).

	Predicted glioma	Predicted meningioma
Actual tumor	905	1
Actual no tumor	1	404

Grad-CAM

To further interpret model decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the ConvNeXt-Tiny model.

The resulting heatmaps, illustrated in Figure 7, clearly highlight the tumor regions that contribute most strongly to the classification output. For *glioma* and *meningioma*, the model focuses accurately on irregular and dense tissue structures within the tumor boundaries, while for *pituitary tumors*, attention is centered around the sellar region. For *no-tumor* cases, activation is diffusely distributed, indicating the absence of focal lesions. These results verify that ConvNeXt-Tiny not only performs with high numerical accuracy but also exhibits interpretability consistent with clinical expectations. The localized activation patterns confirm that the model relies on medically relevant image regions rather than background artifacts, thereby enhancing its trustworthiness and potential for real-world diagnostic support.

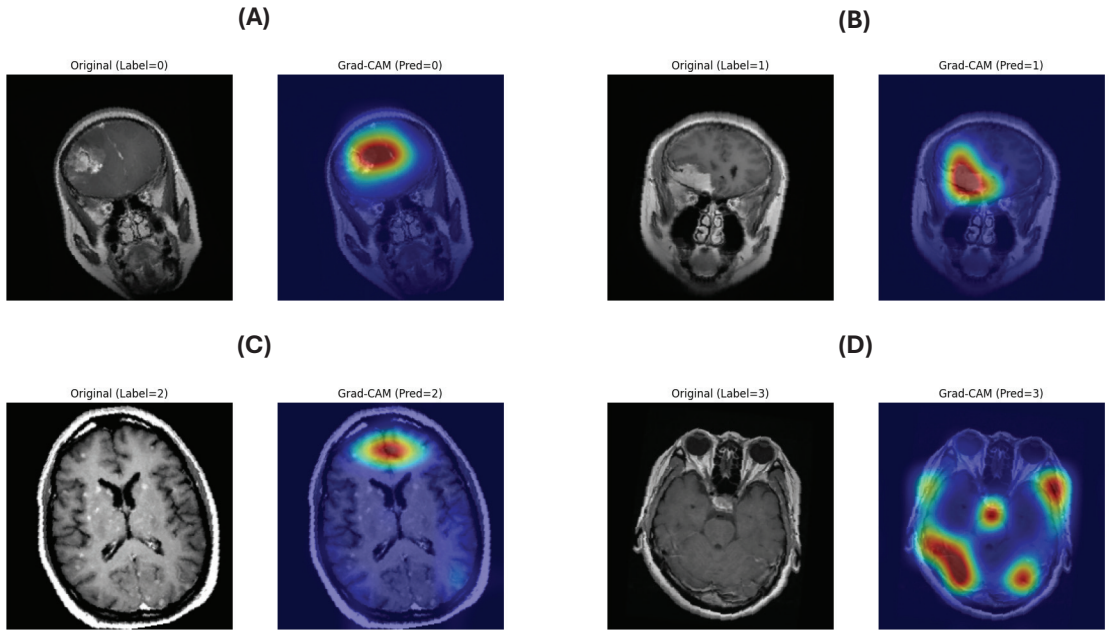

Figure 7. *Grad-CAM visualization of ConvNeXt-Tiny for each class. A: glioma, B: meningioma, C: no tumor, D: pituitary.*

Alternative dataset

To further evaluate the generalization ability of the baseline ConvNeXt-Tiny model, an alternative dataset was employed for testing. This dataset consists of images that were not included in the training process, providing an unbiased assessment of the model's performance on unseen data. The results summarized in Table 11.

Table 11. Performance of ConvNeXt-Tiny in alternative dataset.

Model	Accuracy	Precision	Recall	F1-score
ConvNeXt-Tiny	0.9984	0.9983	0.9982	0.9982

According to Table 11, the ConvNeXt-Tiny model demonstrates strong generalization capability when evaluated on the alternative dataset. The model achieves high performance across all evaluation metrics, with accuracy, precision, recall, and F1-score all exceeding 0.998. Although a slight reduction in accuracy is observed compared to the original training dataset, the overall performance remains consistently high, indicating the robustness and reliability of the proposed model under different data distributions.

Table 12 presents the confusion matrix of the ConvNeXt-Tiny model evaluated on the alternative brain tumor dataset. The results indicate that the model maintains strong generalization performance on unseen MRI scans, with most predictions concentrated along the main diagonal, reflecting a high level of classification accuracy across all four classes: glioma, meningioma, no tumor, and pituitary tumor. Only a small number of misclassifications are observed, mainly between glioma and meningioma, as well as between no-tumor and tumor classes in a few cases. These errors may be attributed to overlapping radiological features in certain MRI samples. Overall, the confusion matrix demonstrates the robustness of ConvNeXt-Tiny in accurately distinguishing no-tumor cases and reliably identifying pituitary tumors.

Table 12. Confusion matrix of the ConvNeXt-Tiny model.

	Predicted glioma	Predicted meningioma	Predicted notumor	Predicted pituitary
Actual glioma	804	1	0	0
Actual meningioma	0	544	0	1
Actual no tumor	0	1	481	1
Actual pituitary	0	0	0	610

Discussion

The comparative evaluation of six convolutional architectures demonstrates that ConvNeXt-Tiny provides a well-balanced trade-off among the evaluated architectures between classification accuracy, computational efficiency, and interpretability. This finding is consistent with prior studies employing ConvNeXt-based models in medical imaging. For example, recent work using pre-trained ConvNeXt on the BraTS 2019 dataset reported competitive or state-of-the-art performance when multi-sequence MRI inputs were available.[23] Although the present study focuses on single-sequence MRI images, ConvNeXt-Tiny achieves comparable performance, indicating its strong feature representation capability under constrained input settings.

While DenseNet169 achieved the highest raw accuracy, ConvNeXt-Tiny exhibited nearly equivalent performance with reduced computational cost and strong generalization on unseen data. These results align with previous MRI-based brain tumor classification studies using transfer learning, which typically report accuracies in the range of 96-98% depending on model complexity and dataset characteristics.[1,6,24] McNemar's test conducted on paired prediction outcomes indicated that performance differences between ConvNeXt-Tiny and other modern CNN architectures were not statistically significant ($p>0.05$), suggesting that ConvNeXt-Tiny should be regarded as competitive rather than strictly superior.

Grad-CAM visualizations showed that ConvNeXt-Tiny consistently focused on clinically relevant tumor regions, supporting findings from prior explainable AI studies in brain MRI analysis.[6] From a practical perspective, ConvNeXt-Tiny occupies a middle ground between heavy and lightweight models, highlighting its suitability for clinical scenarios where both performance and computational efficiency are critical.

Overall, the results position ConvNeXt-Tiny as an interpretable and generalizable baseline for multi-class brain tumor MRI classification. Consistent with prior literature, model selection should consider accuracy alongside efficiency and explainability. Future work should extend validation to multi-center clinical datasets and explore complementary interpretability techniques to further strengthen clinical reliability. To avoid potential metric inflation from the inclusion of the no-tumor class, subtype-specific performance was additionally evaluated, allowing a more realistic assessment of differential diagnosis performance.

Limitations

This study has several limitations. First, all experiments were conducted using publicly available MRI datasets, which may not fully reflect the variability of clinical imaging across different scanners or institutions. Second, the study focused solely on image-level classification without incorporating tumor segmentation, which limits direct clinical interpretability. Third, the interpretability analysis relied exclusively on Grad-CAM, providing only coarse localization rather than precise feature attribution. Finally, although the proposed model demonstrated promising accuracy, further investigation is needed to validate its feasibility for real-time deployment in practical clinical settings. Future work will aim to address these limitations.

Conclusion

This study evaluated the performance of multiple convolutional neural network architectures for multi-class brain tumor classification using MRI images. Among the evaluated models, ConvNeXt-Tiny was employed as the baseline model and demonstrated competitive performance, achieving high accuracy along with balanced precision, recall, and F1-scores, while maintaining a reasonable computational cost. In addition, the model provided interpretable Grad-CAM visualizations that corresponded well with relevant tumor regions, indicating its potential to support more transparent and explainable diagnostic modeling.

Although ConvNeXt-Tiny exhibited strong overall performance, the performance differences among modern architectures such as DenseNet169 and MobileNetV3-Large were relatively small. These observations were further supported by statistical analysis, which revealed no statistically significant differences among the models compared. This suggests that recent CNN-based architectures can achieve comparable classification performance when appropriately trained and optimized.

Overall, the findings highlight ConvNeXt-Tiny as a promising and efficient baseline for comparative research rather than a definitive solution for direct clinical deployment. Future work will focus on further validation in practical usage scenarios, incorporating segmentation-based analysis, and extending interpretability methods beyond Grad-CAM to enhance reliability and transparency in real-world medical imaging applications.

Ethical approval

Ethical approval was not required for this study because all data were obtained from publicly available.

Funding

There was no funding for this work.

Conflict of interest

The authors declare no conflict of interest.

CRediT authorship contribution statement

Nitipon Pongphaw: conceptualization, software, validation, supervisor, writing: original draft, review and editing; **Prommin Buaphan**: data curation, data visualization, writing: review and editing.

Acknowledgements

The author would like to express sincere gratitude to Mr. Keerati Maneesai for valuable advice and assistance in preparing the manuscript.

References

[1] Deepa P, Narain P, Sreena V. Performance analysis of deep transfer learning approaches in detecting and classifying brain tumor from magnetic resonance images. Intelligent Data Analysis. 2023; 27(6):1759-80. doi:10.3233/ida-227321.

[2] Younis A, Li Q, Afzal Z, Adamu M, Kawuwa H, Hussain F, et al. Abnormal brain tumors classification using ResNet50 and its comprehensive evaluation. IEEE Access. 2024; 12: 78843-53. doi:10.1109/access.2024.3403902.

[3] Chen W, Sun P, Zhang Y. A brain tumor diagnosis approach based on deep separable convolutional networks and attention mechanisms. Applied and Computational Engineering. 2024; 67(1): 60-9. doi:10.54254/2755-2721/67/20240629.

[4] Oladimeji O, Ibitoye A. Brain tumor classification using ResNet50-convolutional block attention module. Applied Computing and Informatics. 2023. doi:10.1108/aci-09-2023-0022.

[5] Khan M, Khan A, Alhaisoni M, Alqahtani A, Alsubai S, Alharbi M, et al. Multimodal brain tumor detection and classification using deep saliency map and improved dragonfly optimization algorithm. Int J Imaging Syst Technol. 2023; 33(2): 572-87. doi:10.1002/ima.22831.

[6] Rastogi D, Johri P, Donelli M, Kumar L, Bindewari S, Raghav A, et al. Brain tumor detection and prediction in MRI images utilizing a fine-tuned transfer learning model integrated within deep learning frameworks. Life. 2025; 15(3): 327. doi:10.3390/life15030327.

[7] Li S. Classification of brain tumors based on magnetic resonance imaging using deep learning models. Applied and Computational Engineering. 2024; 40(1): 247-54. doi:10.54254/2755-2721/40/20230660.

[8] Türkoğlu M. Brain tumor detection using a combination of Bayesian optimization based SVM classifier and fine-tuned deep features. Eur J Sci Technol. 2021; 27: 251-258. doi:10.31590/ejosat.963609.

[9] Rehman A, Khan M, Saba T, Mehmood Z, Tariq U, Ayesha N. Microscopic brain tumor detection and classification using 3D CNN and feature selection

architecture. Microsc Res Tech. 2021; 84(1): 133-49. doi:10.1002/jemt.23597.

[10] Murugesan G, Nagendran P, Natarajan J. Advancing brain tumor diagnosis: deep Siamese convolutional neural network as a superior model for MRI classification. Brain-x. 2025; 3(2): e70028. doi:10.1002/brx2.70028.

[11] Kumar P, Bonthu K, Meghana B, Vani K, Chakrabarti P. Multi-class brain tumor classification and segmentation using hybrid deep learning network model. Scalable Comput Pract Exp. 2023; 24(1): 69-80. doi:10.12694/scpe.v24i1.2088.

[12] Ullah M, Khan M, Almujally N, Alhaisoni M, Akram T, Shabaz M. BrainNet: a fusion-assisted novel optimal framework of residual blocks and stacked autoencoders for multimodal brain tumor classification. Sci Rep. 2024; 14(1): 5895. doi:10.1038/s41598-024-56657-3.

[13] Joshi A, Aziz R. Deep learning approach for brain tumor classification using metaheuristic optimization with gene expression data. Int J Imaging Syst Technol. 2024; 34(2): e23007. doi:10.1002/ima.23007.

[14] Pei L, Vidyaratne L, Rahman M, Iftekharuddin K. Context-aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images. Sci Rep. 2020;10(1): 19726. doi:10.1038/s41598-020-74419-9.

[15] Bacak A, Şenel M, Günay O. Convolutional neural network prediction on meningioma and glioma with TensorFlow. Int J Comput Exp Sci Eng. 2023; 9(2): 197-204. doi:10.22399/ijcesen.1306025.

[16] Ullah N, Hassan M, Khan J, Anwar M, Aurangzeb K. Enhancing explainability in brain tumor detection: a novel DeepEBTDNet model with LIME on MRI images. Int J Imaging Syst Technol. 2024; 34(1): e23012. doi:10.1002/ima.23012.

[17] Esmaeili M, Vettukattil R, Banitalebi H, Krogh N, Geitung J. Explainable artificial intelligence for human-machine interaction in brain tumor localization. J Pers Med. 2021; 11(11): 1213. doi:10.3390/jpm11111213.

[18] Goyal D, Sharma H. Brain tumor detection system using neural networks. Int J Commun Inf Technol. 2023; 4(1): 59-63. doi:10.33545/2707661x.2023.v4.i1a.61.

[19] Tonni S, Sheakh M, Tahosin M, Hasan M, Shuva T, Bhuiyan T, et al. A hybrid transfer learning framework for brain tumor diagnosis. Adv Intell Syst. 2025; 7(3): 2400495. doi:10.1002/aisy.202400495.

[20] Alanazi M, Ali M, Hussain S, Zafar A, Mohatram M, Irfan M, et al. Brain tumor/mass classification framework using MRI-based isolated and developed transfer deep-learning model. Sensors. 2022; 22(1): 372. doi:10.3390/s22010372.

[21] Nickparvar M. Brain tumor MRI dataset [Internet]. Kaggle; 2021 [cited 2025 Jan 3]. Available from: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[22] Rostami A. Labeled MRI Brain Tumor Dataset Computer Vision Model [Internet]. Roboflow Universe; 2024 [cited 2025 Jan 3]. Available from: https://universe.roboflow.com/ali-rostami/labeled-mri-brain-tumor-dataset

[23] Mehmood Y, Bajwa UI. Brain tumor grade classification using the ConvNeXt architecture. Digit Health. 2024; 10: 20552076241284920. doi:10.1177/20552076241284920.

[24] Aamir M, Namoun A, Munir S, Aljohani N, Alanazi MH, Alsahafi Y, et al. Brain tumor detection and classification using an optimized convolutional neural network. Diagnostics. 2024; 14(16): 1714. doi:10.3390/diagnostics14161714.

Supplymentary

1. Dataset source

Dataset 1

The brain MRI dataset used in this study was compiled from multiple publicly available sources to form a four-class classification problem consisting of glioma, meningioma, pituitary tumor, and no-tumor categories. Images for the glioma, meningioma, and pituitary tumor classes were primarily obtained from the figshare dataset, which contains T1-weighted contrast-enhanced MRI scans acquired using clinical scanners from Nanfang Hospital (Guangzhou, China) and Tianjin Medical University General Hospital between 2005 and 2010. This dataset provides slice-level annotations for three tumor types and has been widely adopted in prior brain tumor classification studies. The no-tumor class was sourced from the Br35H dataset, which consists of brain MRI images from healthy subjects and is commonly used for binary and multi-class brain tumor classification tasks. Although the SARTAJ dataset was initially considered as an additional source for tumor images, the glioma subclass was excluded due to observed label inconsistencies, as indicated by prior studies and confirmed through our own experimental results. To ensure label reliability and reduce scanner-induced bias, glioma samples were therefore retained exclusively from the figshare dataset. The final curated dataset contains 7,023 MRI images with clearly defined class origins, enabling consistent multi-class training and evaluation.

Alternative Dataset

This project constructed a labeled brain MRI dataset for multi-class tumor analysis, covering four categories: pituitary tumor, meningioma tumor, glioma tumor, and no tumor. The dataset consists of a total of 2,443 MRI images, which were systematically divided into training, validation, and test subsets comprising 1,695, 502, and 246 images, respectively; however, in this study, all available images were additionally merged and used as an external test set to evaluate the generalization capability of the proposed models. All images are magnetic resonance imaging (MRI) scans, and each sample was annotated by medical experts following a standardized labeling protocol. The annotations include tumor presence and tumor type, with additional information on tumor location when applicable. This dataset is designed to support the development and evaluation of machine learning and deep learning models for automated brain tumor classification, with potential applications in assisting radiologists during clinical diagnosis and facilitating research toward improved diagnostic tools and treatment planning.

2. Loss curve

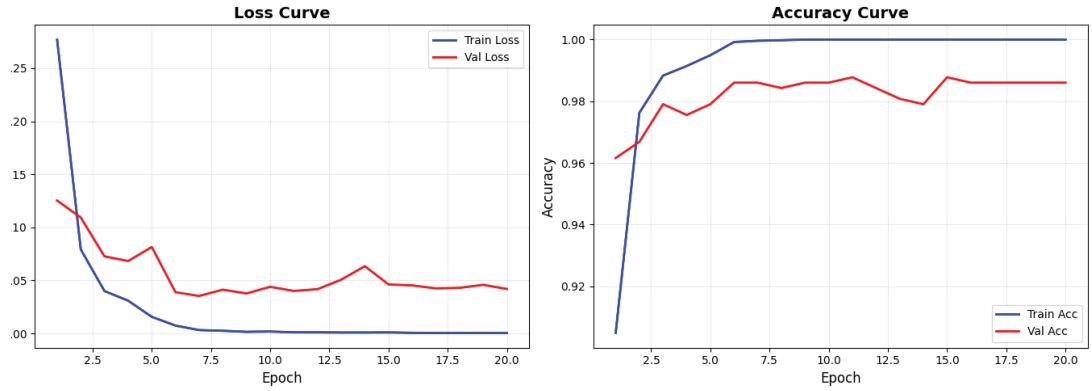

Figure S1. *Training and validation performance curves of the CNN+DenseNet169.*

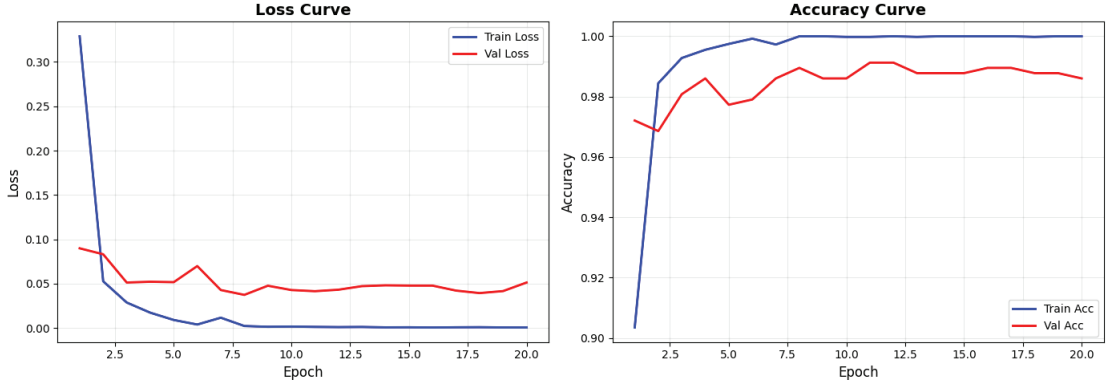

Figure S2. *Training and validation performance curves of the Xception.*

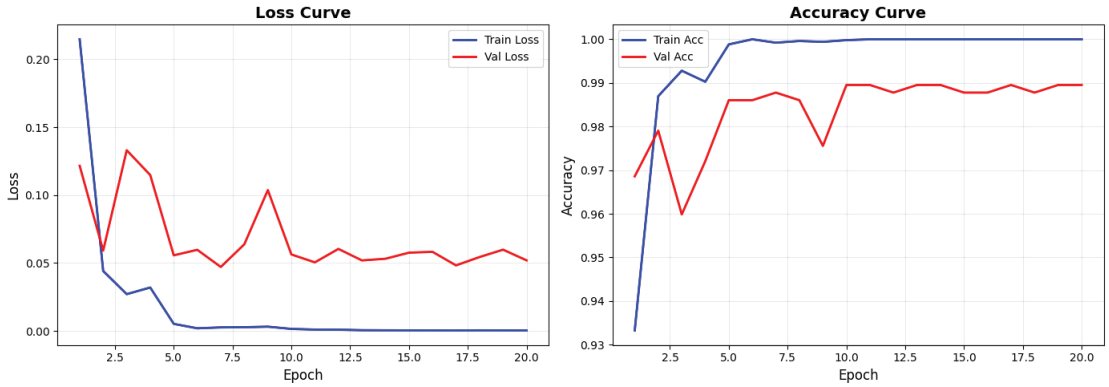

Figure S3. *Training and validation performance curves of the DenseNet169.*

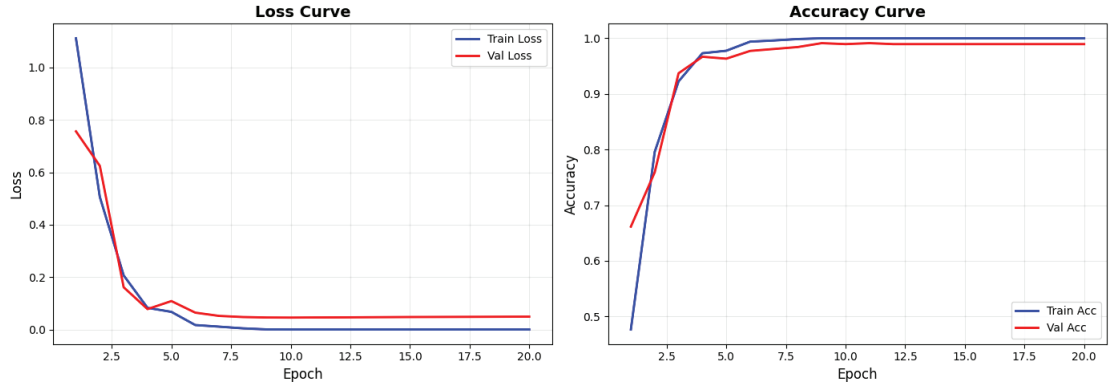

Figure S4. *Training and validation performance curves of the ConvNextTiny.*

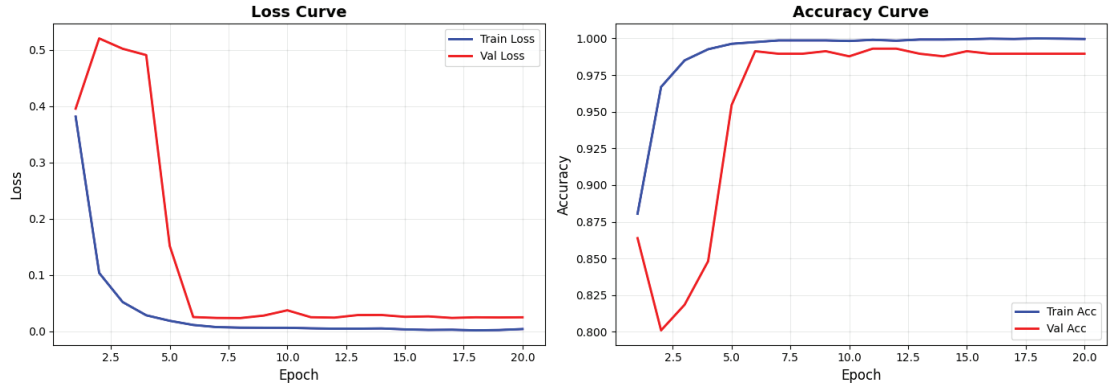

Figure S5. *Training and validation performance curves of the MobileNetV3.*

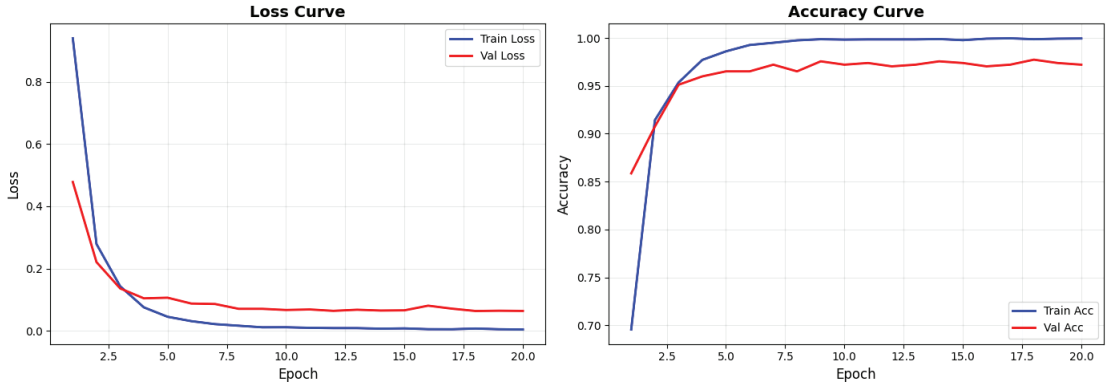

Figure S6. *Training and validation performance curves of the ResNet50.*

3. Confusion matrix

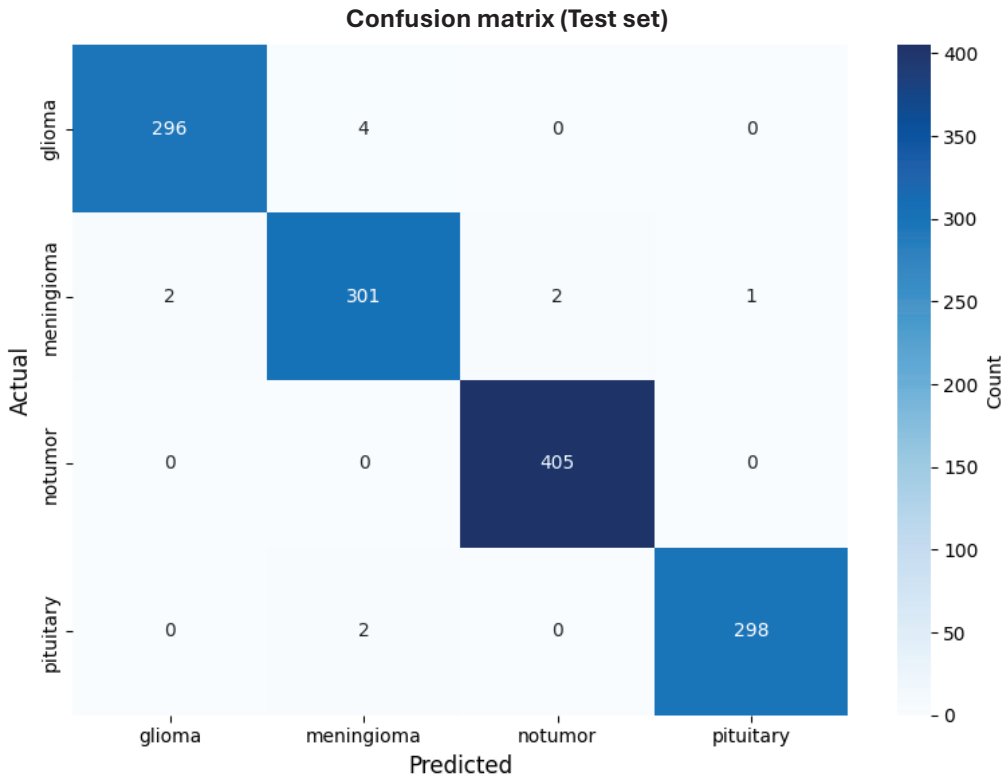

Figure S7. *Confusion matrix of the CNN+DenseNet169 model.*

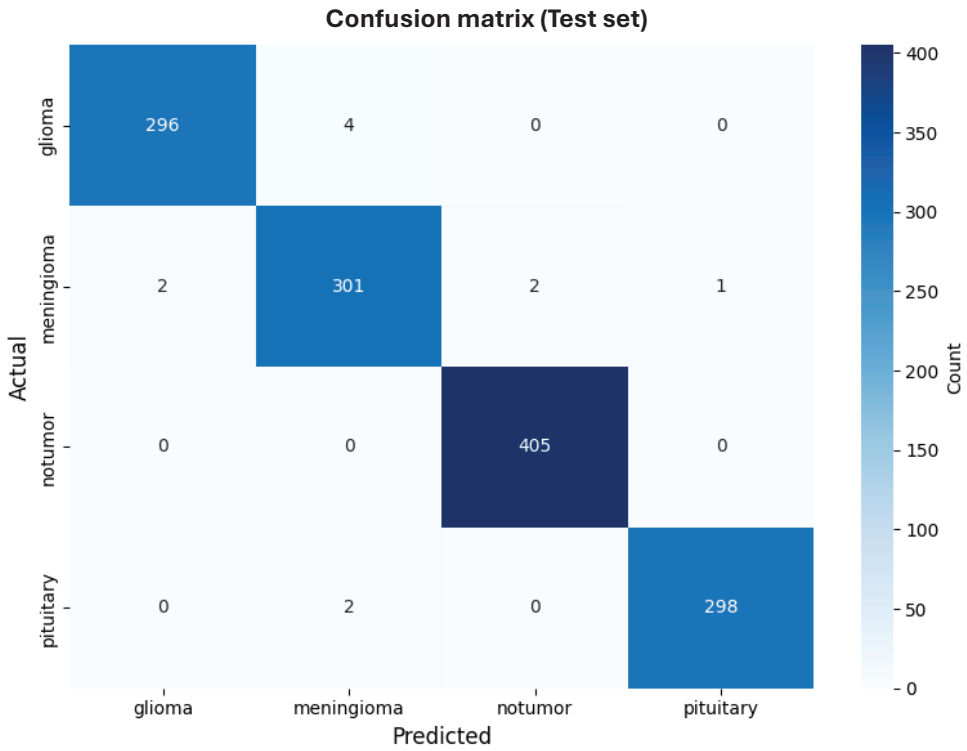

Figure S8. *Confusion matrix of the Xception model.*

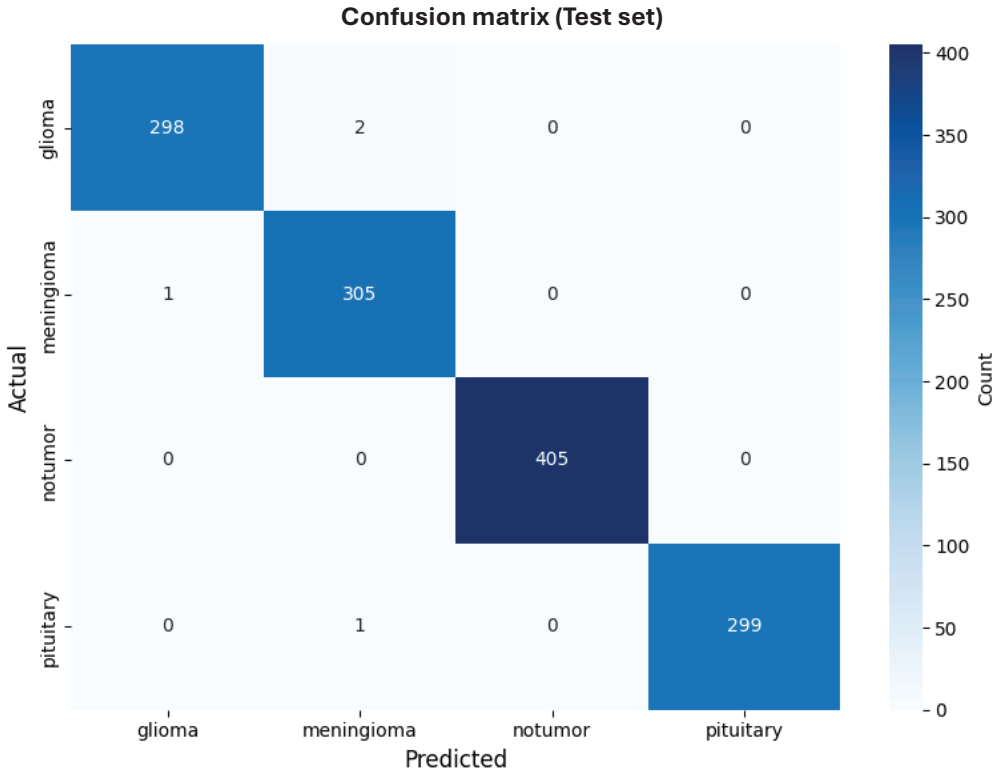

Figure S9. *Confusion matrix of the DenseNet169 model.*

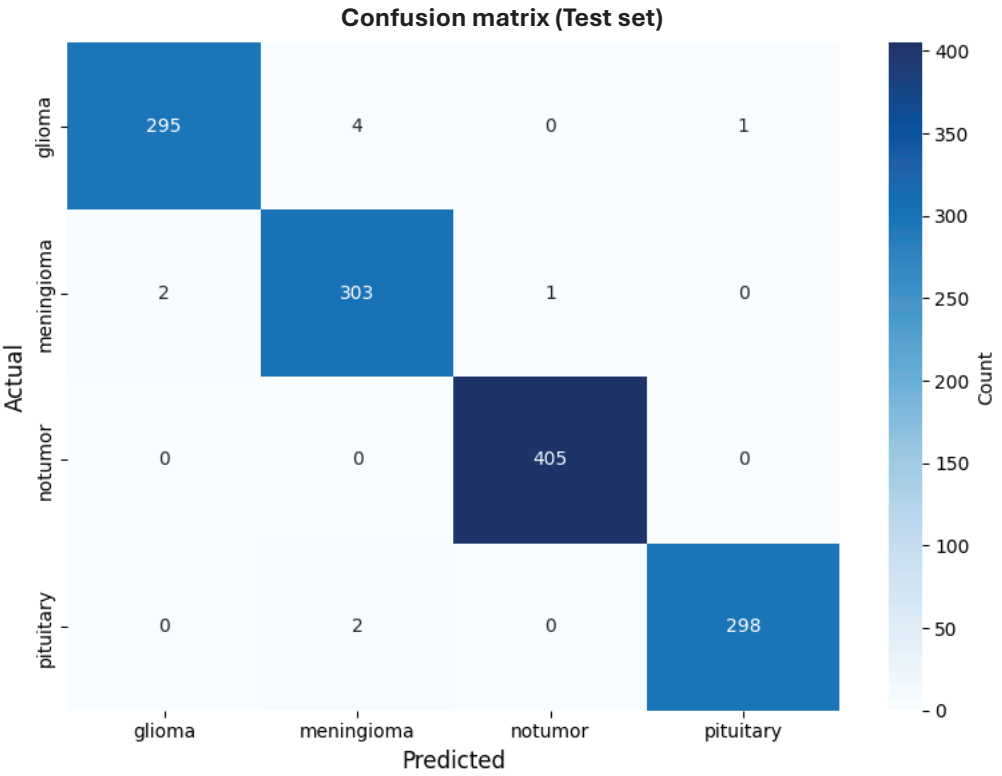

Figure S10. *Confusion matrix of the ConvNextTiny model.*

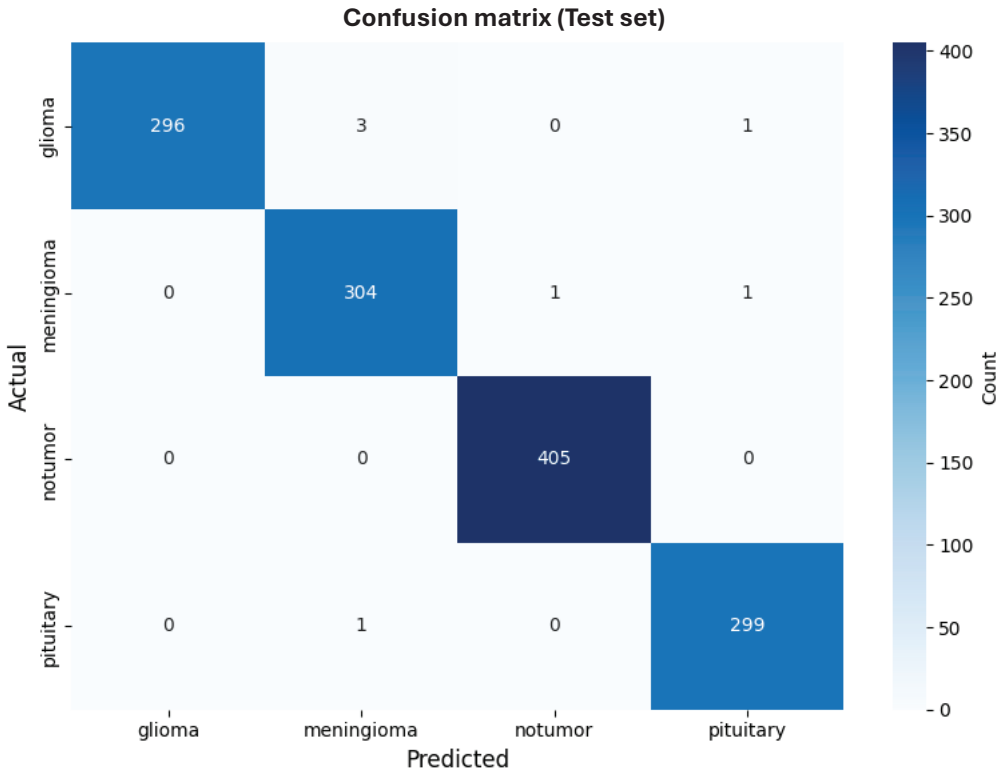

Figure S11. *Confusion matrix of the MobileNetV3 model.*

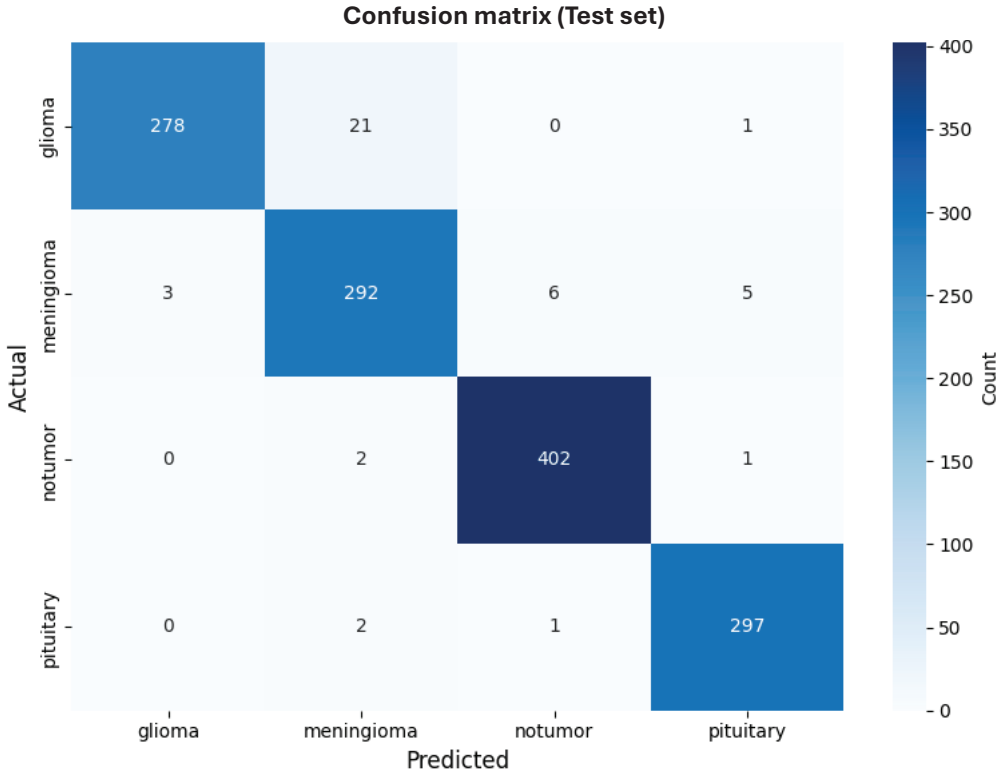

Figure S12. *Confusion matrix of the ResNet50 model.*

4. Subclass tumor

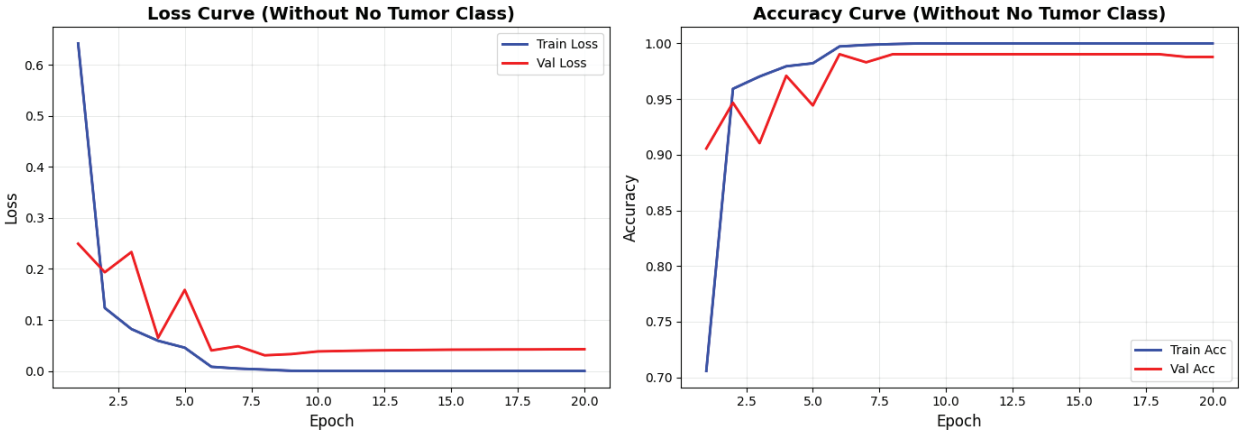

Figure S13. *Training and validation performance curves of the ConvNextTiny.*

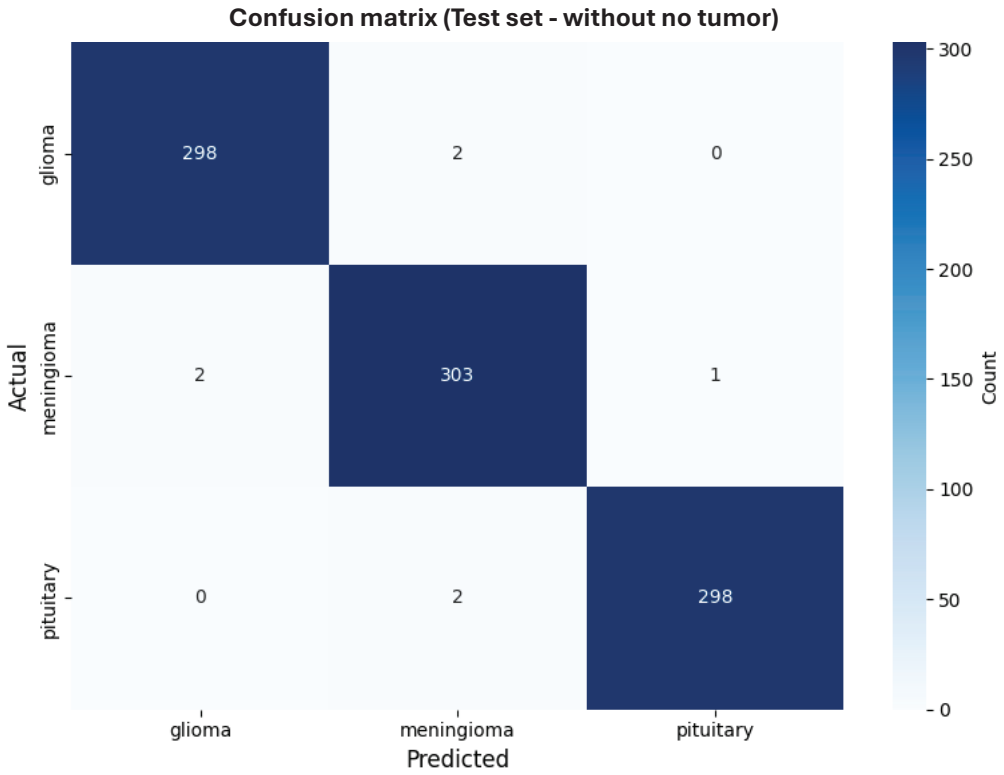

Figure S14. *Confusion matrix of the ConvNextTiny model.*

5. Tumor vs no tumor

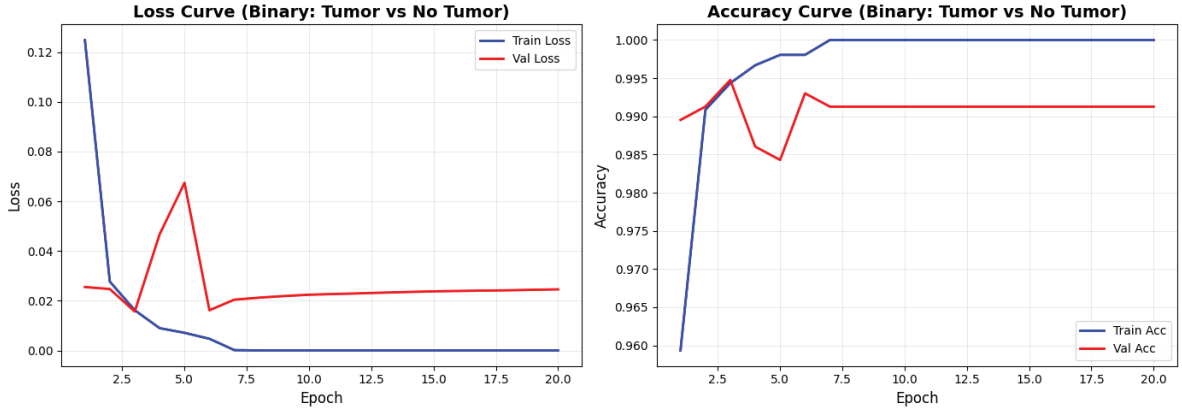

Figure S15. *Training and validation performance curves of the ConvNextTiny.*

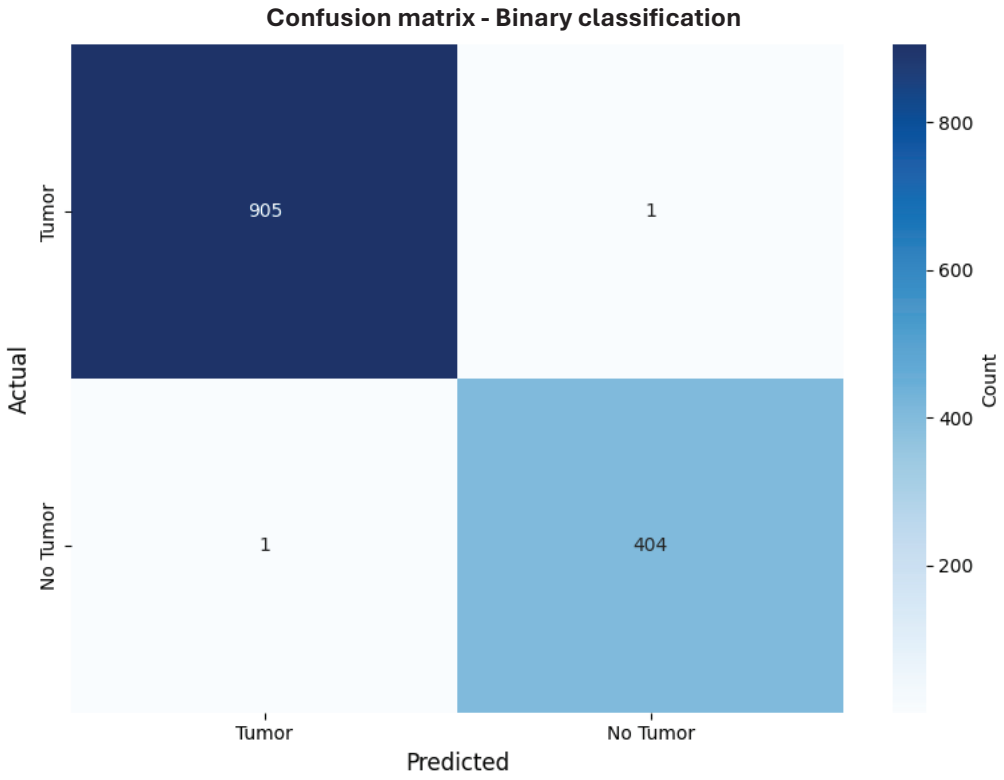

Figure S16. *Confusion matrix of the ConvNextTiny model.*

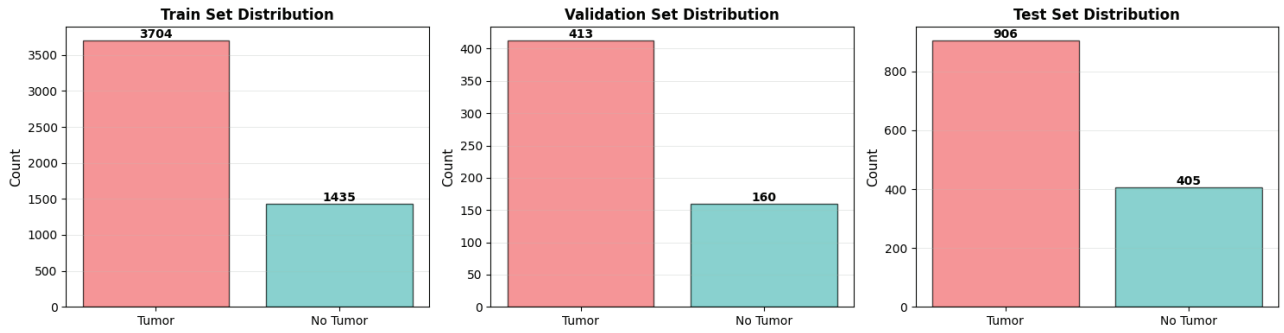

Figure S17. *Set distributions.*