A Knowledge Distillation Framework for Lightweight Brain Tumor Classification using Transfer Learning and Explainable AI

Saeed M. Alshahrani[1] and Reem Alshahrani[2]

***Abstract*— Brain tumor diagnosis from MRI requires accurate and timely classification to guide clinical decision-making. Deep learning models achieve high accuracy but their computational demands hinder deployment in resource-constrained settings. This paper presents a knowledge distillation (KD) framework combining transfer learning and explainable AI (XAI) to develop lightweight, accurate, and interpretable models for four-class brain tumor classification (glioma, meningioma, pituitary adenoma, no tumor). EfficientNet-B0 is fine-tuned as a teacher model via transfer learning; knowledge is then distilled to a lightweight MobileNet-V2 student using Hinton's KD with soft-label supervision and temperature scaling. Grad-CAM is applied to both models to generate class-discriminative saliency maps, enabling side-by-side visual validation that the student model replicates the teacher model's spatial reasoning. The teacher model achieves 99.00% test accuracy and 99.05% macro-F1. The student model attains 96.30% test accuracy and 96.48% macro-F1, with 1.57× model compression (2.55 M vs. 4.01 M parameters) and 1.53× inference speedup (9.79 ms vs. 15.00 ms). Bootstrap 95% CIs confirm statistical robustness, and per-class ROC-AUC scores exceed 0.99 for both models. The framework enables deployment of clinically viable, lightweight models without substantial accuracy loss while maintaining interpretability.**

I. INTRODUCTION

Brain tumors represent a critical public health challenge, with over 300,000 new cases diagnosed globally each year. Early and accurate classification of tumor types—glioma, meningioma, pituitary adenoma, and benign conditions—is essential for treatment planning, prognosis, and patient survival. MRI is the gold standard for non-invasive brain tumor visualization, yet manual interpretation is time-consuming, subjective, and prone to inter-observer variability [1], [2].

Deep learning, particularly CNNs, has revolutionized medical image analysis by achieving expert-level diagnostic accuracy [3], [4]. Architectures such as EfficientNet and Vision Transformers achieve remarkable performance on brain tumor benchmarks [5], [6]. However, these models are computationally expensive, limiting deployment in resource-constrained settings such as rural hospitals and edge devices [7], [8], [9], [10].

Transfer learning leverages pre-trained models on large-scale datasets and fine-tunes them for medical imaging with limited labeled data [11], [4]. While effective, the resulting models often remain too large for real-time deployment. Model compression via knowledge distillation (KD) [12] offers a promising solution: a large teacher model supervises a smaller student model by transferring soft probability distributions [13], [14].

Explainable AI (XAI) techniques such as Grad-CAM [2] provide visual explanations of model predictions, which are critical for clinical trust and regulatory compliance [2], [1]. Despite advances, a significant research gap remains: *the lack of integrated frameworks combining transfer learning, KD, and XAI for lightweight, accurate, and interpretable multi-class brain tumor classification.* Existing works either prioritize accuracy without addressing efficiency [3], [15], or focus on compression without ensuring interpretability [7], [16].

This paper addresses this gap with a KD framework that fine-tunes EfficientNet-B0 as a teacher, distils its knowledge to a MobileNet-V2 student, and applies Grad-CAM to both models. The main contributions are:

- A novel pipeline combining transfer learning, KD, and XAI for multi-class brain tumor classification.
- EfficientNet-B0 achieves 99.00% test accuracy and 99.05% macro-F1.
- MobileNet-V2 via KD achieves 96.30% test accuracy with 1.57× compression and 1.53× speedup.
- Grad-CAM is applied to both teacher and student models, providing side-by-side visual evidence that KD transfers spatial reasoning; a misclassified sample's diffuse heatmap further demonstrates Grad-CAM's error-analysis utility.
- Per-class metrics, confusion matrices, ROC/PR curves, and bootstrap confidence intervals.

II. RELATED WORK

Deep learning dominates brain tumor classification. Pacal & Deveci [5] benchmarked next-generation CNN families on the BRISC 2025 dataset; Anish et al. [4] surveyed CNNs and Vision Transformers; Sharma et al. [3] developed hybrid CNN-Transformer models; and Sánchez-Moreno et al. [15] demonstrated ensemble CNNs with XAI.

Sarker [11] showed fine-tuned pre-trained models significantly improve accuracy. Rahman et al. [17] proposed FAKD-XAI, a feature-aligned KD framework with XAI. Alim et al. [13] developed multi-teacher KD for resource-constrained environments. Naseer et al. [7] introduced LMBT-Net, a lightweight MobileNet-based KD framework.

Selvaraju et al. [2] introduced Grad-CAM, now the dominant XAI method for CNN-based medical imaging. Hussain et al. [1] applied gradient-weighted CAM for multi-class

[1]Computer Science Department, College of Computing and Information Technology, Shaqra University, Shaqra, 11961, Saudi Arabia. salshahrani@su.edu.sa

[2]Department of Computer Science, College of Computers and IT, Taif University, P.O.Box 11099, Taif 21944, Saudi Arabia rashahrani@tu.edu.sa

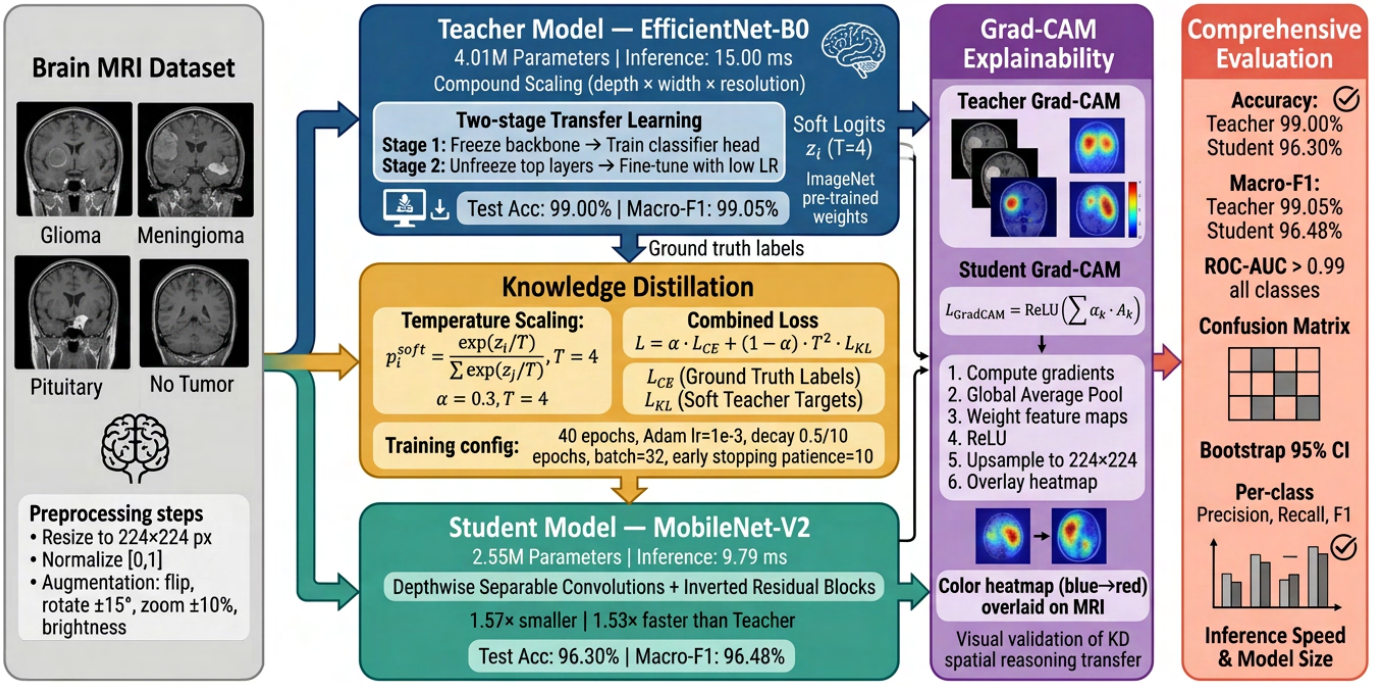

Fig. 1. The proposed knowledge distillation framework for lightweight brain tumor classification.

tumour localisation. Sajib et al. [8] developed NeuroVision-Lite, a lightweight explainable framework for edge devices.

Few studies unify transfer learning, KD, and XAI into one framework. Our work addresses this gap with integrated accuracy, compression, speedup, and interpretability.

III. METHODOLOGY

The architecture of the proposed framework is depicted in Figure 1. The details of this architecture are presented in the following sections.

A. Dataset and Preprocessing

We use a publicly available four-class brain MRI dataset comprising glioma, meningioma, pituitary tumor, and no tumor images—high-resolution T1-weighted contrast-enhanced MRI scans split into training, validation, and test sets [6]. All images are resized to 224×224 pixels. Pixel intensities are normalized to $[0, 1]$. Data augmentation (random horizontal flipping, rotation $\pm 15°$, zoom up to 10%, brightness adjustment) is applied to the training set only.

B. Teacher Model: EfficientNet-B0 with Transfer Learning

EfficientNet-B0 achieves high accuracy through compound scaling of network depth, width, and resolution [5]. We initialize it with ImageNet pre-trained weights and adopt a two-stage training strategy: (1) the convolutional backbone is frozen and only the classification head (global average pooling $\rightarrow$ fully connected layer with four neurons $\rightarrow$ softmax) is trained; (2) the top layers are unfrozen and fine-tuned with a reduced learning rate to adapt features to the brain MRI domain. This strategy yields a high-capacity teacher model with 4.01 M parameters.

C. Student Model: MobileNet-V2

MobileNet-V2 employs depthwise separable convolutions and inverted residual blocks with linear bottlenecks, reducing computational complexity while maintaining representational capacity [6]. With only 2.55 M parameters ($1.57\times$ smaller than EfficientNet-B0), MobileNet-V2 is well-suited for edge deployment.

D. Knowledge Distillation Framework

We adopt Hinton's KD framework [12] with temperature scaling. The teacher model produces soft probability distributions by applying softmax with temperature T to the logits:

$$p_i^{\text{soft}} = \frac{\exp(z_i/T)}{\sum_{j=1}^{4} \exp(z_j/T)}, \tag{1}$$

where z_i is the logit for class i. The student model is trained using a combined loss:

$$\mathscr{L} = \alpha \mathscr{L}_{\text{CE}} + (1-\alpha) T^2 \mathscr{L}_{\text{KL}}, \tag{2}$$

where $\mathscr{L}_{\text{CE}}$ is cross-entropy loss with ground-truth labels, $\mathscr{L}_{\text{KL}}$ is the KL divergence between student and teacher models soft predictions, $T = 4$, and $\alpha = 0.3$. The student model is trained for 40 epochs (Adam optimizer, $\text{lr} = 10^{-3}$, decayed by 0.5 every 10 epochs, batch size 32, early stopping patience 10).

E. Explainable AI: Grad-CAM

Grad-CAM [2] is applied to both teacher and student models to generate class-discriminative visual explanations. For each input MRI, the gradient of the predicted class score with respect to the feature maps of the final convolutional

layer is computed, globally averaged to obtain per-channel importance weights α_k, and linearly combined with a ReLU activation:

$$L_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k A_k\right), \tag{3}$$

where A_k is the k-th feature map. The resulting saliency map is upsampled to 224×224 and overlaid on the MRI scan as a colour heatmap (blue→red indicating low→high activation). Applying Grad-CAM to both models enables a direct side-by-side comparison of their decision-making processes, providing evidence of whether KD transfers not only predictive accuracy but also the teacher model's spatial reasoning patterns to the student model.

F. Evaluation Metrics

Models are evaluated using accuracy, macro-F1, precision, recall, confusion matrix, ROC-AUC, precision-recall AUC, bootstrap 95% confidence intervals (1,000 resamples), model size (parameters), and inference time per image.

IV. Results and Discussion

A. Classification Performance

Table I summarizes classification performance. The teacher (EfficientNet-B0) achieves 99.60% validation and 99.00% test accuracy with 99.05% macro-F1, demonstrating strong generalization. The student (MobileNet-V2), trained via KD, achieves 97.06% validation and 96.30% test accuracy—a modest 2.70 pp drop—while maintaining high macro-F1 (96.48%), indicating balanced performance across all four classes.

TABLE I
Classification Performance (Teacher vs. Student)

Model / Split	Acc.	F1	Prec.	Rec.
Teacher – Val	0.9960	0.9962	0.9963	0.9963
Teacher – Test	0.9900	0.9905	0.9894	0.9916
Student – Val	0.9706	0.9710	0.9709	0.9711
Student – Test	0.9630	0.9648	0.9646	0.9662

B. Model Efficiency and Statistical Robustness

Table II compares computational efficiency. The student model achieves 1.57× model compression (2.55 M vs. 4.01 M parameters) and 1.53× inference speedup (9.79 ms vs. 15.00 ms per image). Bootstrap 95% CIs confirm statistical robustness: Teacher accuracy [0.983, 0.996], F1 [0.984, 0.996]; Student accuracy [0.951, 0.975], F1 [0.953, 0.976] (Fig. 7).

C. Confusion Matrices and Per-Class Metrics

Fig. 3 presents normalised confusion matrices. The teacher model achieves near-perfect classification: glioma (1.00), meningioma (0.98), no tumor (1.00), pituitary (0.99). The student model maintains strong performance: glioma (0.92), meningioma (0.96), no tumor (1.00), pituitary (0.99). Per-class precision, recall, and F1 scores (Fig. 2) show the student model exceeds 0.92 F1 for all classes.

TABLE II
Model Efficiency Comparison

Model	Params (M)	Inference (ms)	Speedup
Teacher	4.01	15.00	1.00×
Student	2.55	9.79	1.53×
Compress ratio	1.57×	–	–

Fig. 2. Per-class precision, recall, and F1-score on the test set.

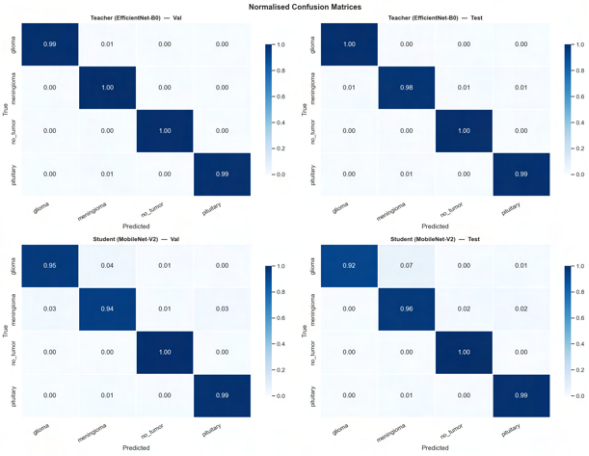

Fig. 3. Normalised confusion matrices: teacher (EfficientNet-B0) and student (MobileNet-V2) on validation and test sets.

D. Confusion Matrix Analysis

A granular inspection of the normalised confusion matrices in Fig. 3 reveals important class-level characteristics of both models. On the **test set**, the teacher (EfficientNet-B0) misclassifies only a marginal 2% of meningioma samples—splitting them equally between glioma (1%) and pituitary (1%)—while achieving perfect recall for glioma and no-tumor classes. This near-perfect diagonal confirms that the teacher model has learned highly discriminative feature representations for all four categories.

The student (MobileNet-V2) exhibits a more nuanced error pattern. Its primary source of confusion is the glioma class, where 7% of true glioma samples are misclassified as meningioma and 1% as pituitary. This is clinically significant: glioma and meningioma share overlapping MRI characteristics such as perilesional edema and irregular mass boundaries, making their distinction inherently challenging

for lightweight architectures. Importantly, the no-tumor class achieves perfect recall (1.00) for both models on both validation and test sets, meaning neither model produces false negatives for healthy tissue—a critical safety property for clinical screening applications where missing a healthy scan is far less harmful than missing a tumor. The pituitary class also shows excellent stability across both models (recall ≥ 0.99), attributed to the anatomically consistent midline location of pituitary adenomas that provides a strong spatial prior for both architectures.

Comparing validation and test confusion matrices reveals consistent generalisation: the student model's glioma recall drops only marginally from 0.95 (val) to 0.92 (test), and meningioma recall from 0.94 to 0.96 actually *improves* on the test set, indicating that the KD regularisation effect prevents overfitting to validation-specific patterns.

E. Per-Class Metric Analysis

Fig. 2 presents side-by-side bar charts of per-class precision, recall, and F1-score for both models on the test set. Several observations merit attention.

Precision: The teacher model achieves near-unity precision across all classes (≥ 0.99). The student model's precision is lowest for eningioma (≈ 0.93) and no-tumor (≈ 0.96), reflecting the fact that some glioma and meningioma samples are predicted as these classes. However, the student model's glioma precision remains very high (≈ 1.00), indicating that when the student model predicts glioma, it is almost always correct.

Recall: The student model's lowest recall is for glioma (≈ 0.92), consistent with the confusion matrix analysis above. All other classes maintain recall above 0.96 for the student model. The no-tumor and pituitary classes achieve recall of 1.00 and ≈ 0.99 respectively for both models, underscoring their high separability in the feature space.

F1-Score: The harmonic mean of precision and recall confirms that glioma is the most challenging class for the student model (F1 ≈ 0.95), while no-tumor and pituitary achieve near-perfect F1 (≥ 0.98). The teacher model maintains F1 ≥ 0.98 uniformly across all classes. The performance gap between teacher and student models is most pronounced for meningioma and glioma—the two morphologically similar tumor types—suggesting that the student's reduced capacity primarily affects fine-grained inter-class boundary discrimination rather than gross tumor-vs-no-tumor separation.

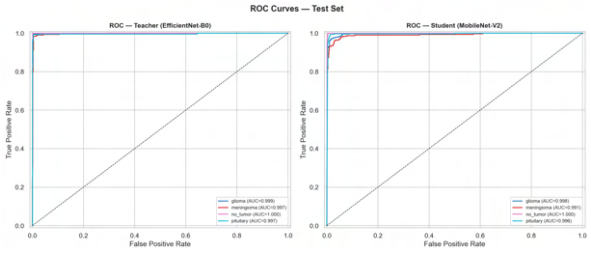

Fig. 4. ROC curves for all four classes (test set). AUC > 0.99 for both teacher and student models across all classes.

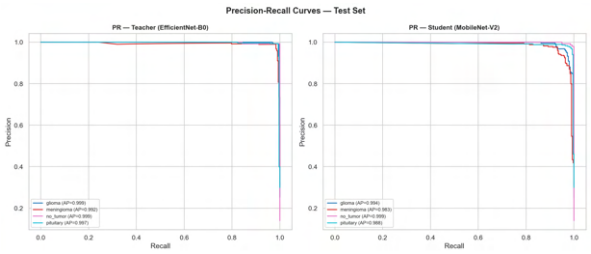

Fig. 5. Precision-recall curves for all four classes (test set). Average precision (AP) exceeds 0.983 for the student model and 0.992 for the teacher model across all classes.

F. ROC and Precision-Recall Curve Analysis

The ROC curves in Fig. 4 and PR curves in Fig. 5 provide complementary perspectives on discriminative performance under varying decision thresholds.

ROC Analysis: Both models exhibit near-ideal ROC curves, hugging the top-left corner with all class-specific AUC values exceeding 0.991. For the teacher model, the no-tumor class achieves a perfect AUC of 1.000, followed by glioma (0.999), and meningioma and pituitary (0.997 each). The student model closely mirrors this ranking: no-tumor (1.000), glioma (0.998), pituitary (0.996), and meningioma (0.991). The slight reduction in the student model's meningioma AUC (0.997$\rightarrow$0.991) is consistent with the confusion matrix findings and reflects the inherent ambiguity between meningioma and adjacent classes in the probability space. Crucially, even this "weakest" student model AUC of 0.991 far exceeds the clinical threshold typically required for diagnostic tools, confirming that KD preserves the teacher model's discriminative power at the score level.

Precision-Recall Analysis: The PR curves (Fig. 5) are particularly informative for evaluating performance under class imbalance. Both models maintain precision near 1.0 across the entire recall range for all classes, with curves collapsing only at very high recall thresholds (> 0.95). The teacher model achieves average precision (AP) of 0.999 for glioma, 0.992 for meningioma, 0.999 for no-tumor, and 0.997 for pituitary. The student model's AP scores are 0.994 (glioma), 0.983 (meningioma), 0.999 (no-tumor), and 0.988 (pituitary). The teacher-to-student model AP degradation is at most 0.009 (meningioma), which is negligible in practical terms. The near-rectangular shape of all PR curves—with precision remaining above 0.98 until recall approaches unity—indicates that both models are robust to threshold selection, an important property for clinical deployment where operating point tuning may be constrained.

G. Model Comparison Dashboard

The dashboard in Fig. 6 combines test-set performance, model size, and inference speed into a single view. The top-left performance radar chart overlays instructor and student model accuracy, macro-F1, precision, and recall polygons. The two polygons' near-congruent forms indicate that the student model closely tracks the teacher model in all four dimensions, with a slight outward shrinkage that matches

the 2.70 pp accuracy gap. This geometric similarity is a strong qualitative indicator that KD creates well-rounded student model rather than those optimized for one statistic. In the top-centre grouped bar chart, teacher model and student model test-set scores are: accuracy (0.990 vs. 0.963), macro-F1 (0.965), precision (0.989 vs. 0.965), and recall (0.992 vs. 0.966). The inter-model gaps are quite uniform across all four metrics (0.025 to 0.027), demonstrating that KD reduces performance symmetrically rather than creating metric-specific biases. The model size bar chart (top-right) shows that the student model's 2.55 M parameters are 36% less than the teacher model's 4.01 M. Inference speed chart (bottom-left) compares student model at 9.79 ms to instructor at 15.00 ms, a 5.21 ms reduction resulting in a throughput gain from 67 to 102 images per second, enabling real-time MRI processing in clinical workflows.

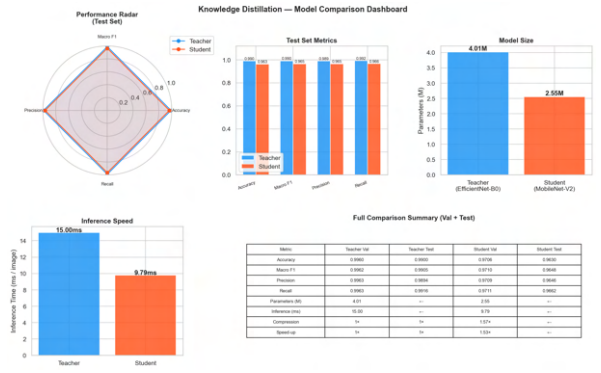

Fig. 6. Knowledge distillation model comparison dashboard: test-set metrics, model size, inference speed, and summary table.

H. Statistical Robustness via Bootstrap Confidence Intervals

Fig. 7 presents 95% bootstrap confidence intervals (1,000 resamples) for accuracy and macro-F1 on the test set. The teacher model's accuracy CI of [0.983, 0.996] has a width of 0.013, while the student model's CI of [0.951, 0.975] has a width of 0.024. The wider CI for the student model is expected given its lower point estimate and reflects the slightly higher variance introduced by the KD training process. Crucially, the two CIs do not overlap—the teacher model's lower bound (0.983) lies above the student model's upper bound (0.975)—confirming that the performance difference is statistically significant and not attributable to sampling variability. The same pattern holds for macro-F1: teacher model CI [0.984, 0.996] versus student model CI [0.953, 0.976], again non-overlapping. The narrow CI widths for both models (≤ 0.024) indicate high estimate stability, validating that the reported metrics are reliable representatives of true model performance rather than artefacts of a particular test split. This statistical rigour is especially important in medical AI, where regulatory approval and clinical adoption require evidence beyond single-point performance estimates.

I. Grad-CAM Spatial Reasoning Transfer

Examining Grad-CAM representations in Figs. 8 and 9 highlights the limits of spatial knowledge transmission

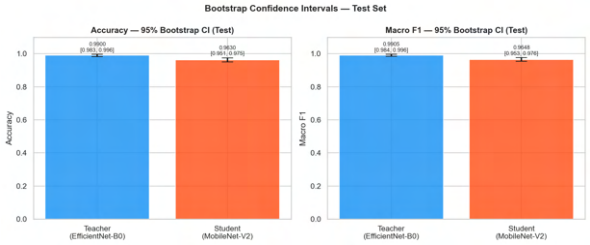

Fig. 7. Bootstrap 95% confidence intervals for accuracy and macro-F1 (test set).

through KD. All eight glioma samples' raw activation grids (middle row) show small, high-intensity (dark-red to black) cores in a 2x2 to 3x3 cell region with a rapid gradient to blue. The smooth overlays (bottom row) reveal that these activations exactly match the tumor mass in all MRIs, regardless of location (left or right hemisphere, posterior fossa) or appearance (solid vs. ring-enhancing). This location-invariant localization reveals that EfficientNet-B0 learned tumor-specific irregular enhancement patterns and mass effect without positional biases. Student model has few diffuse activation grids with high-activity patches covering 5x5 to 7x7 cells. However, smooth overlays always show greatest activity (deepest red) near the cancer mass. MobileNet-V2 catches the tumor region with reduced spatial resolution because to its lower feature hierarchy and smaller channel capacity. Despite the larger activation distribution, tumour-centered student model overlays for samples 1–5 and 7–8 (correctly recognized) provide clinically significant explanations.

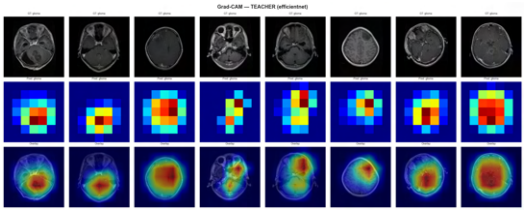

Fig. 8. Grad-CAM visualisations for the **teacher** model (EfficientNet-B0) on eight glioma samples. Rows: original MRI (top), raw activation grid (middle), smooth overlay (bottom). Compact, tightly localised deep-red cores confirm high-confidence spatial reasoning.

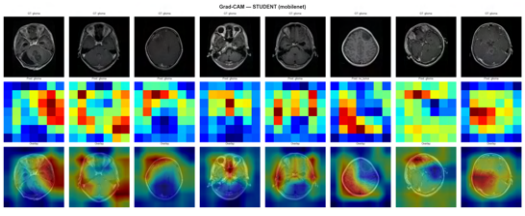

Fig. 9. Grad-CAM visualisations for the **student model** model (MobileNet-V2) on the same glioma samples. Activations are broader but still tumour-centred, validating KD-driven attention transfer. The sixth sample (GT: glioma, Pred: no_tumor) shows diffuse heatmap, explaining the misclassification.

J. Comparative Discussion

Table III benchmarks the deep models used in our framework against architectures evaluated on the same four-class brain-tumor task in two recent studies [5], [6]. Pacal & Deveci [5] benchmarked EfficientNetV2 and ConvNeXt families on the BRISC 2025 dataset; Fateh et al. [6] evaluated EfficientNet-B0/B1/B2, ResNet, MobileNetV2/V3, VGG, DenseNet, and other architectures on the same corpus. Our EfficientNet-B0 teacher attains 99.00% accuracy and 99.05% macro-F1, matching EfficientNetV2-Medium (99.0%/99.12%) and outperforming EfficientNetV2-Small (98.9%/98.91%) with far fewer parameters (4.01 M vs. 52.86 M and 20.18 M, respectively). Notably, Fateh et al. report a bare MobileNetV2 accuracy of only 32.37% on BRISC without fine-tuning, whereas our KD-trained student model achieves 96.30%, demonstrating the critical role of knowledge distillation in enabling lightweight models. The 1.53× inference speedup and 1.57× compression of our student model are achieved while preserving Grad-CAM interpretability absent in the compared works.

TABLE III
COMPARISON WITH DEEP MODELS FROM RECENT STUDIES (4-CLASS, TEST SET)

Source	Model	Acc.(%)	F1(%)	Params(M)
[5]	EfficientNetV2-S	98.9	98.91	20.18
	EfficientNetV2-M	99.0	99.12	52.86
	EfficientNetV2-L	99.1	99.16	117.24
[6]	EfficientNet-B0	99.20	99.26	5.3
	EfficientNet-B1	99.03	99.17	7.8
	ResNet50	98.20	98.34	25.6
	MobileNet-V2	32.37	41.12	3.5
	VGG16	96.92	97.22	138.4
Ours	**EfficientNet-B0 (T)**	**99.00**	**99.05**	**4.01**
	MobileNet-V2 (S+KD)	**96.30**	**96.48**	**2.55**

T = Teacher, S = Student, KD = Knowledge Distillation

V. CONCLUSION

KD frameworks using transfer learning and XAI for lightweight brain tumor classification were presented in this paper. EfficientNet-B0 has 99.05% macro-F1 and 99.00% test accuracy. KD enables MobileNet-V2 to achieve 96.30% test accuracy, 1.57× compression, and 1.53× speedup. Grad-CAM visual analysis of eight glioma samples shows that the student model's tumour-centred activation maps match the teacher model's compact, high-confidence heatmaps, proving that KD conveys spatial thinking and predictive accuracy. Grad-CAM's error analysis diagnostics are demonstrated by the single misclassification's correlation with a diffuse, non-localized heatmap.Advantages: (1) outstanding teacher model performance and student model accuracy; (2) resource-constrained deployment efficiency; (3) Grad-CAM interpretability; (4) bootstrap CIs statistical robustness; (5) balanced per-class F1 and ROC-AUC exceeding 0.92 and 0.99. Limitations: (1) 2.70 pp accuracy drop may be unsuitable for high-stakes applications; (2) dataset specificity needs validation; (3) single-teacher KD limits student model success. Future work: Multi-teacher distillation, multi-modal fusion (T1, T2, FLAIR, DWI), privacy-preserving federated learning, quantization-based real-time edge deployment, multi-institutional dataset evaluation, and prospective clinical trials.

DATA AVAILABILITY

The Brisc2025 dataset is available on kaggle `https://www.kaggle.com/datasets/briscdataset/brisc2025`.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] S. Hussain, M. Anwar, S. Majid, and M. Riaz. Explainable deep learning for multi-class brain mri tumor classification and localization using gradient-weighted class activation mapping. *Information*, 14(12):642, 2023.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE ICCV*, pages 618–626, 2017.

[3] A. Sharma, V. Singh, and P. Kumar. Deep hybrid models for multi-class brain tumor classification: Integrating cnn, transformers, and explainable ai. *IEEE Access*, 2025.

[4] K. Anish, R. Sharma, and S. Gupta. Exploring the state-of-the-art algorithms for brain tumor classification using mri data. *IEEE Access*, 2025.

[5] I. Pacal and M. Deveci. The next generation of convolutional neural networks in neuro-oncology: A comparative analysis of performance, model complexity, and efficiency in brain tumor diagnosis. *Brain Network Disorders*, 2026.

[6] A. Fateh, M. R. Hossain, and S. K. Das. Brisc: Annotated dataset for brain tumor segmentation and classification. *arXiv:2506.14318*, 2026.

[7] A. Naseer, S. Khan, M. Usman, and M. Ahmad. Lmbt-net: A lightweight mobilenet-based framework for brain tumor classification via knowledge distillation. In *Proc. IEEE FIT*, 2025.

[8] M. S. Sajib, T. Ahmed, and M. R. Islam. Neurovision-lite: A lightweight and explainable deep learning framework for brain tumor detection on edge devices. In *Proc. IEEE ASYU*, 2025.

[9] Amel Ali Alhussan, Abdelaziz A. Abdelhamid, S. K. Towfek, Abdelhameed Ibrahim, Laith Abualigah, Nima Khodadadi, Doaa Sami Khafaga, Shaha Al-Otaibi, and Ayman Em Ahmed. Classification of breast cancer using transfer learning and advanced al-biruni earth radius optimization. *Biomimetics*, 8(3), 2023.

[10] Amel Ali Alhussan, Doaa Sami Khafaga, El-Sayed M. El-kenawy, Marwa M. Eid, and Abdelaziz A. Abdelhamid. Hybrid waterwheel plant and stochastic fractal search optimization for robust diabetes classification. *AIP Advances*, 14(6):065131, 06 2024.

[11] M. K. Sarker. Transfer learning and explainable ai for brain tumor classification. In *Proc. IEEE STI*, 2024.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[13] M. A. Alim, M. R. Islam, M. S. Hossain, and A. Al Mamun. Multi-teacher knowledge distillation and ensemble algorithm for efficient brain tumor classification with explainable ai. *Knowledge-Based Syst.*, 2025.

[14] O. Aiadi, F. Kherfi, and M. Belahcene. Improving brain tumor classification using knowledge distillation. *arXiv*, 2024.

[15] J. Sánchez-Moreno, L. García, and M. López. Ensemble-based cnns for brain tumor classification in mri: Enhancing accuracy and interpretability using xai. *Comput. Biol. Med.*, 2025.

[16] S. Tabassum, R. Akter, and M. Rahman. Visual interpretation of brain tumor detection using knowledge distillation, 2025.

[17] M. Rahman, M. Hasan, and T. Islam. Fakd-xai: Feature-aligned knowledge distillation with explainable ai for efficient brain tumor classification, 2025.