FAKD-XAI: Feature-Aligned Knowledge Distillation with Explainable AI for Efficient Brain Tumor Classification

Md. Abdur Rahman[1], Sabik Aftahee[2], Md. Ashiqur Rahman[3], and Lamim Zakir Pronay[4]

[1] Southeast University, Dhaka, Bangladesh
2021200000025@seu.edu.bd
[2] Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh
sabikaftahee.official@gmail.com
[3] Southeast University, Dhaka, Bangladesh
ashiqur.rahman@seu.edu.bd
[4] National Institute of Technology Andhra Pradesh,India
pronayfarab03@gmail.com

Abstract. Accurate and efficient classification of brain tumors by magnetic resonance imaging (MRI) scans is essential for clinical follow-up and treatment planning. However, in deep learning models, computational costs are often a significant barrier to practical application. This paper presents Feature-Aligned Knowledge Distillation with XAI (FAKD-XAI), a novel framework that classifies and rationally interprets brain tumors in an efficient manner. FAKD-XAI combines logit-level Knowledge Distillation with an adaptive intermediate feature-level distillation from ResNet-50 (Teacher Model) to a lightweight MobileNetV3-Large (Student Model) to facilitate learning between complex and simple models. Our alignment module featuring a 1×1 convolution layer was able to overcome the architectural divergences of the student model and enabled the efficient use of stratified feature transfer at different levels of the hierarchy. FAKD-XAI integrates Local Interpretable Model-agnostic Explanations (LIME), which enhances the understanding of the workings behind model predictions, leading to promoting trust from the clinicians. FAKD-XAI achieved an accuracy of 99.47% on the Brain Tumor MRI dataset while maintaining high computational efficiency, with an average inference time of 5.25 ms per image. This makes it highly suitable for practical, clinical deployment. The use of Explainable AI (XAI) confirms that the model focuses on pertinent tumor areas, suggesting FAKD-XAI's usefulness as a reliable diagnostic aid. All code is available on GitHub: https://github.com/borhanitrash/FAKD-XAI

Keywords: Brain Tumor Classification · Magnetic Resonance Imaging (MRI) · Feature-Aligned Knowledge Distillation (FAKD) · Explainable Artificial Intelligence (XAI) · Lightweight Convolutional Neural Networks · Computational Efficiency.

1 Introduction

Brain tumors are among the most life-threatening forms of cancer, contributing to significant mortality and morbidity worldwide. In 2022, it was reported that nearly 700,000 people were living with a brain tumor in the United States alone, with approximately 84,000 new cases diagnosed annually [1]. Early and accurate classification of brain tumors is critical for treatment planning and improving survival rates.

Magnetic Resonance Imaging (MRI) remains the gold standard for brain tumor diagnosis due to its high-resolution imaging capabilities [2]. However, manual diagnosis is time-consuming, subjective, and requires expert radiologists, which makes automated classification methods highly desirable. In recent years, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in medical imaging applications, including brain tumor classification [3–5].

Despite their impressive performance, deep learning models often involve complex architectures with millions of parameters, leading to high computational costs and memory requirements [6]. This makes them unsuitable for deployment in resource-constrained environments such as mobile healthcare applications or rural clinics. Knowledge Distillation (KD) has emerged as a promising technique to mitigate these limitations by transferring knowledge from a large teacher model to a smaller student model without significant performance degradation [7].

Traditional KD methods focus mainly on matching the output logits of the teacher and student networks [7]. However, recent research highlights that incorporating intermediate feature representations during distillation can lead to substantial improvements in student performance, especially in tasks requiring fine-grained pattern recognition, such as brain tumor classification [8, 9].

In this paper, we propose a novel framework, Feature-Aligned Knowledge Distillation with XAI (FAKD-XAI), to enhance the accuracy, efficiency, and interpretability of brain tumor classification. The key contributions of our proposed method are as follows:

- Presented a new framework called **Feature-Aligned Knowledge Distillation with XAI (FAKD-XAI)** that enables efficient knowledge transfer from a deep teacher *(ResNet-50)* [10] to an effective student model *MobileNetV3-Large* [11] by combining *logit-level* distillation with adaptively aligned intermediate *feature-level* distillation.
- Integrated Explainable AI (XAI) using *LIME* [12] with the distilled student model, offering visual representations of the model's decision-making process to improve clinical trust and transparency.
- Achieved 99.47% accuracy on the Brain Tumor MRI dataset with lower computational cost, outperforming baseline models, suitable for real-world deployment.

2 Literature Review

Several approaches have been made to detect and classify brain tumors using deep learning and Knowledge Distillation (KD) methods. Jiang et al. [13] used KD to train a five-layer student CNN from a DenseNet121 teacher model. The student model achieved an impressive 97.48% accuracy, sometimes surpassing the teacher model. The authors used average map visualizations across the convolutional layers of the student model. The authors utilized the same dataset, which includes four tumor classes, as described in the dataset section. Gohari et al. [14] approached this problem using a combination of federated learning and KD to classify brain tumors into three classes. They used the VGGNet16 teacher model trained on MRI data to distill the knowledge into the student model, helping it achieve 94.38% accuracy. Anantathanavit et al. [15] trained ResNet18 as the teacher model on a small dataset of 357 MRI images. The student model achieved 98.10% accuracy using fewer resources, producing results comparable to larger models like VGG. Kanchanamala et al. [16] proposed a hybrid QDCNN-DMN model that enhanced MRI images using logarithmic transformations to classify tumors into three types: GD-ET, ED, and NCR/NET. Zarenia et al. [17] introduced a framework for automated brain tumor classification and segmentation using a multiscale deformable attention module (MS-DAM). The MS-DAM model achieved over 96.5% classification accuracy and performed tumor segmentation to enhance diagnostic precision. Guan et al. [18] proposed a framework for automated brain tumor classification using low-quality MRI images, achieving a classification accuracy of 98.04% on a public dataset. Chaitanya and Satpathy et al. [19] proposed a knowledge-distilled ResNeXt-50 model that preprocesses MRI images and classifies brain tumors using transfer learning and KD. The student model achieved 95.3% accuracy, outperforming models like VGG16.

3 Materials and method

In this section, we present our proposed, Feature-Aligned Knowledge Distillation with XAI (FAKD-XAI) framework for brain tumor classification. The proposed method transfers knowledge from a bigger teacher network to a more effective student network, which is especially tailored for medical image analysis tasks. It's done by combining multi-level knowledge transfer through feature alignment and distillation.

3.1 Dataset Description

This study used the Brain Tumor MRI dataset from Kaggle [20]. This dataset is a combination of the SARTAJ, Br35H, and figshare datasets. It included 7,023 brain MRI images in four different classes: pituitary tumor (1,757 images), meningioma (1,645 images), glioma (1,621 images), and no tumor (2,000 images). The dataset was pre-divided into training (5,712 images) and testing (1,311 images) sets. The images displayed differences in intensity, contrast, noise levels,

and anatomical characteristics, emulating real clinical environments. These natural differences enhance model development with practical clinical applicability and enhance generalization capacities for brain tumor classification tasks.

3.2 Dataset Preprocessing

Initially, we split the training set, allocating 80% of the data for model training and 20% for validation, as this dataset lacked a designated validation set. We converted the images to RGB format to ensure compatibility with pre-trained models. Images were downsized to 224×224 pixels using bilinear interpolation and normalized with parameters (mean = [0.5, 0.5, 0.5], std = [0.5, 0.5, 0.5]). We applied data augmentation techniques, including random horizontal flipping ($p = 0.5$) and random rotation ($\pm 15°$), to improve model generalization while maintaining anatomical integrity. A tailored `BrainMRIDataset` class was created with thorough error management, ensuring uniform class-to-index relationships. The data loading process utilized PyTorch's `DataLoader`, implementing batch sizes of 32 for training and validation, and 16 for testing, while enabling parallel processing to improve training efficiency.

3.3 Overview of the Distillation Framework

Knowledge distillation (KD) has become a prominent technique for compressing deep neural networks while maintaining performance. As initially formulated by Hinton *et al.* [7], KD transfers knowledge from a teacher model to a student model by training a smaller student model to replicate the output of a larger teacher model. The conventional KD loss mostly focuses on the soft output predictions. However, recent studies [8,9] showed that aligning intermediate feature representations captures detailed information, which can be significantly beneficial to the student model, especially in the medical image analysis domain where pattern recognition is crucial.

As illustrated in Fig. 1, our FAKD-XAI framework extends the conventional KD approach by integrating feature-level information transfer with logit-level distillation while maintaining direct supervision from ground truth labels.

3.4 Network Architecture

Teacher Network: We chose the ResNet-50 architecture, pre-trained on ImageNet, as our teacher network. With its 50 convolutional layers and residual connections, ResNet-50 has demonstrated remarkable performance across various computer vision tasks, including medical image classification. In our approach, the final fully connected layer was modified to output predictions for our specific brain tumor classification task, encompassing four classes: pituitary, meningioma, glioma, and no tumor.

The depth and complexity of ResNet-50 enable it to learn rich hierarchical features, making it an ideal teacher model. However, its computational requirements (approximately 25.6 million parameters) can limit its deployment in resource-constrained medical environments.

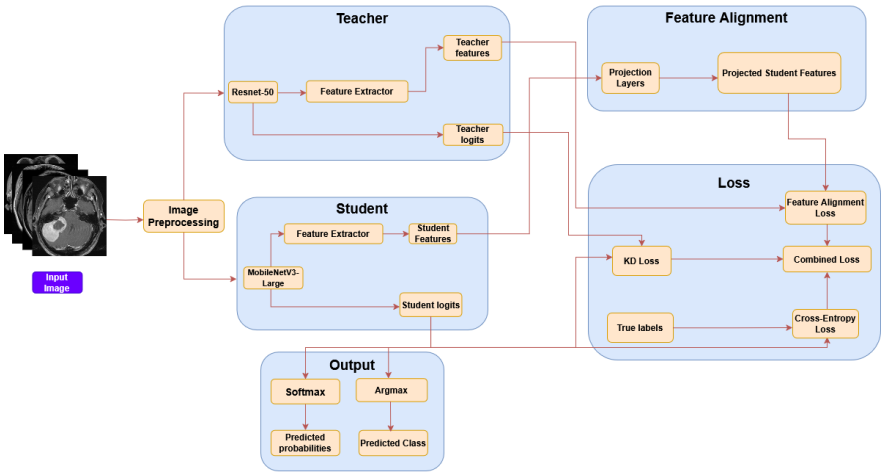

Fig. 1. FAKD-XAI Frameworks

Student Network: MobileNetV3-Large, which uses effective depthwise separable convolutions and squeeze-and-excitation blocks to achieve good performance with much fewer parameters, is used as our student network. We modify the final classifier layer to fit our four-class classification task. MobileNetV3-Large is suitable for deployment in memory-constrained situations or mobile healthcare applications because it has about 7.5 times fewer parameters than ResNet-50 while still performing competitively.

3.5 Intermediate Feature Extraction

In order to capture richer, hierarchical feature representations beyond the final output logits, the hybrid distillation framework heavily relies on knowledge transfer from intermediate layers. To enable this, a `FeatureExtractor` wrapper module was created. During the forward pass, this wrapper uses PyTorch's `register_forward_hook` to intercept and capture activation maps from specific intermediary layers of the teacher and student models.

The output of the `layer3` residual block is used to extract features for the ResNet-50 teacher model, providing a typical mid-to-high level embedding. For the MobileNetV3-Large student architecture, activations are retrieved at the `features.16` stage (with the entire `features` module serving as a fallback), which precedes the final pooling and classification segments. These selection decisions help mitigate architectural discrepancies by aligning the semantic depth captured by both models.

3.6 Feature Alignment Module

Directly comparing feature maps from the teacher and student networks is problematic when their spatial dimensions (height and width) and channel depths

are not the same. To overcome this problem, a two-stage alignment process is used prior to the computation of the feature-based distillation loss.

For spatial alignment, the student's feature maps (`student_features`) are first upsampled to the spatial resolution of the teacher's feature maps (`teacher_features`). This resizing is performed using bilinear interpolation (i.e., `F.interpolate` with `mode='bilinear'` and `align_corners=False`).

The number of channels may still differ even after matching spatial dimensions. Channel alignment is introduced as a solution to this issue. It is a lightweight projection layer, implemented as a single 1×1 convolution (`torch.nn.Conv2d`), placed on top of the upsampled student features. This convolutional layer adjusts the channel count to match that of the teacher's features. Its parameters are optimized jointly with the student model during training. These two modules ensure that student and teacher feature maps are directly comparable by first aligning spatial dimensions and then equalizing channel depths. This allows for an efficient feature-based similarity loss computation.

3.7 Feature-Aligned Knowledge Distillation

Our FAKD-XAI framework incorporates three complementary components in the loss function to effectively transfer knowledge.

Cross-Entropy Loss: To ensure direct supervision from the ground labels, the traditional cross-entropy loss is utilized [21]. The formula of Cross-Entropy Loss is shown in equation 1

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\big(p_{i,c}\big) \tag{1}$$

where $y_{i,c}$ is the ground truth label, $p_{i,c}$ is the student's predicted probability, N is the batch size, and C is the number of classes.

Logit-Level Knowledge Distillation: In accordance with Hinton et al. [7], we employ temperature-scaled softening of the logits to transfer the relationship information between different classes from teacher to student. The formula of KD Loss is shown in equation 2

$$\mathcal{L}_{KD} = D_{KL}\big(\text{softmax}(z^s/T),\, \text{softmax}(z^t/T)\big) \;\cdot\; T^2 \tag{2}$$

where z^s and z^t denote the logits from the student and teacher networks respectively, T is the temperature parameter controlling the softness of the probability distribution, and D_{KL} represents the Kullback–Leibler divergence.

Higher temperature values smooth the probability distribution, emphasizing the relationships between different classes rather than just the correct class. We set $T = 3.0$ based on empirical evaluation.

Feature Alignment Loss: To capture and transfer the rich intermediate representations from the teacher to the student, we introduce a spatial feature alignment mechanism. Unlike prior approaches that require matching feature dimensions, we employ a 1×1 convolutional projection layer to adapt the student's feature map dimensions to match those of the teacher. We measure the discrepancy between these feature maps using mean squared error, as shown in equation 3

$$\mathcal{L}_{FA} = \left\| P_\theta\left(F^s_{\text{up}}\right) \; - \; F^t \right\|_2^2 \tag{3}$$

where F^s is the student's feature map, F^t is the teacher's feature map, F^s_{up} is the student's feature map upsampled to match the teacher's spatial dimensions using bilinear interpolation, and P_θ is a learnable projection layer with parameters θ.

We extract features from the penultimate layer (`layer3`) of ResNet-50 for the teacher and from the last feature block (`features.16`) of MobileNetV3-Large for the student. This specific selection is based on semantic similarity and optimization of the knowledge transfer process.

Combined Loss Function: The overall loss function combines the three components with weighting factors. The formula of Combined Loss Function is shown in equation 4

$$\mathcal{L}_{total} = (1 - \alpha) \cdot \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{KD} + \beta \cdot \mathcal{L}_{FA} \tag{4}$$

where α controls the balance between cross-entropy and KD loss, and β determines the importance of feature alignment loss. We empirically set $\alpha = 0.5$ and $\beta = 1.0$ to optimize performance.

3.8 Explainable Artificial Intelligence (XAI)

We used Explainable AI (XAI) techniques to improve interpretability and transparency. We employed Local Interpretable Model-agnostic Explanations (LIME) for its ability to explain any black-box model and its intuitive, superpixel-based visualizations. While we acknowledge its potential for explanation instability, its local fidelity provides valuable, case-specific insights into the model's reasoning. By changing the input image (using superpixels) and building a simple, interpretable model based on these changes, LIME explains predictions locally, weighted by their proximity to the original instance. This process identifies the input areas of the superpixels that are most relevant for the decision making of the model. We integrated LIME with our PyTorch model and its specific preprocessing procedures using the `lime` library to create a custom prediction function (`predict_fn_lime`).

To facilitate visualization, we produced explanations for a varied collection of test images, emphasizing the superpixels that contribute positively (supporting evidence for the predicted class) and negatively (contradicting evidence) using

color-coded boundaries superimposed on the original image. This qualitative examination ascertains whether the model emphasizes clinically pertinent image aspects, thus fostering confidence in its predictions. We configured LIME with `num_samples=1000` perturbations and aimed to identify the top three contributing features for each explanation.

4 Results and Discussion

This section describes the experimental setup, evaluation metrics, and thorough result analysis we got from our proposed framework. We contrast its performance with the baseline student model and many state-of-the-art techniques. We also applied XAI methods to interpret the predictions of the model and for transparency.

4.1 Environment Setup

Experiments were run on the Kaggle platform utilizing cloud computing resources optimized for deep learning applications. The environment consisted of an Intel Xeon CPU with 2 cores and a system RAM of 29 GB. For GPU acceleration, an NVIDIA Tesla P100 was used, equipped with 16 GB of VRAM.

4.2 Hyperparameter Tuning

The student network's training and the feature alignment projection layer were guided by a set of hyperparameters that were carefully chosen, as shown in Table 1. We employed the AdamW optimizer [22], well known for its efficient regularization using decoupled weight decay. The learning rate was controlled by a cosine annealing schedule [23] with a maximum of 15 epochs to promote convergence. The training used a batch size of 32 and was enhanced by Automatic Mixed Precision (AMP) with gradient scaling to optimize computing resources and maintain numerical stability.

To mitigate overfitting, in addition to optimizer-level weight decay and data augmentation, we employed early stopping based on validation accuracy, stopping training if no improvement was detected for five successive epochs. The model checkpoint with the best validation accuracy was preserved for final testing. Key parameters influencing the knowledge distillation loss, including the temperature T and loss weighting factors α and β, were established by empirical tuning to equilibrate learning from ground truth, teacher logits, and feature alignment. The settings of the teacher network were maintained constant during this operation.

4.3 Experimental Results

Training Behavior Analysis: Figure 2 shows the training and validation performance curves for the proposed FAKD-XAI framework. The left panel depicts

Table 1. Key Model Hyperparameters

Hyperparameter	Value
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
Weight Decay	1×10^{-5}
Learning Rate Schedule	Cosine Annealing
Batch Size	32
Distillation Parameters	
Temperature (T)	3.0
KD Loss Weight (α)	0.5
Feature Loss Weight (β)	1.0

the loss curves of the training and validation, and the right panel shows the accuracy curves of the training and validation throughout 15 epochs. A consistent decline in both training and validation losses indicates efficient learning and convergence. Likewise, the training and validation accuracy rise slowly and plateau. This indicates that the model generalizes well to the unseen validation data without notable overfitting. The dashed vertical line marks Epoch 10, where the saved model checkpoint utilized for the next evaluation was created with the highest validation accuracy. The entire training process was completed in 6.16 minutes. Each epoch demonstrated high efficiency, taking an average of 24.6 seconds to complete.

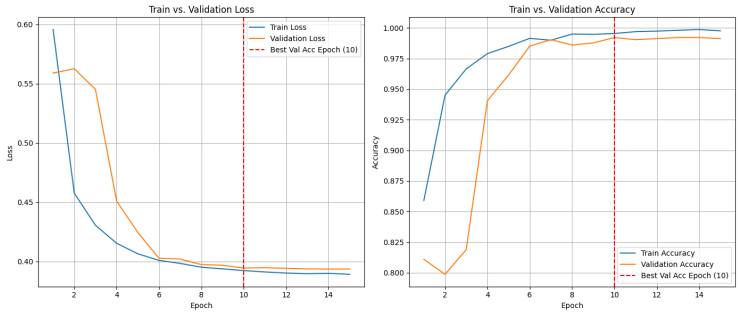

Fig. 2. Training and validation performance curves for the FAKD-XAI framework.

Quantitative Evaluation of FAKD-XAI: The performance of the best FAKD-XAI model on the independent test set was evaluated using the standard evaluation metrics. The overall classification report is provided in Table 2.

In the test set, the proposed FAKD-XAI model obtained an exceptional total accuracy of 99.47%. The precision, recall, and F1-scores for every class are

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
glioma	99.00	100.00	99.00	99.50
meningioma	99.02	98.70	99.02	98.86
notumor	100.00	99.51	100.00	99.75
pituitary	99.67	99.67	99.67	99.67
macro avg	99.42	99.47	99.42	99.44
weighted avg	**99.47**	99.47	99.42	99.47

Table 2. Classification performance metrics by class for proposed FAKD-XAI framework.

extraordinarily high, mostly above 0.99, suggesting strong performance across all tumor types and the 'notumor' class.

Further visualizing the performance of the model, the confusion matrix in Figure 3 shows the accuracy for each class. The strong diagonal entries confirm the great accuracy for each class. Misclassifications are few; only three glioma cases were misclassified as meningioma, and one meningioma case was misclassified as pituitary. The model accurately identified every "notumor" instance.

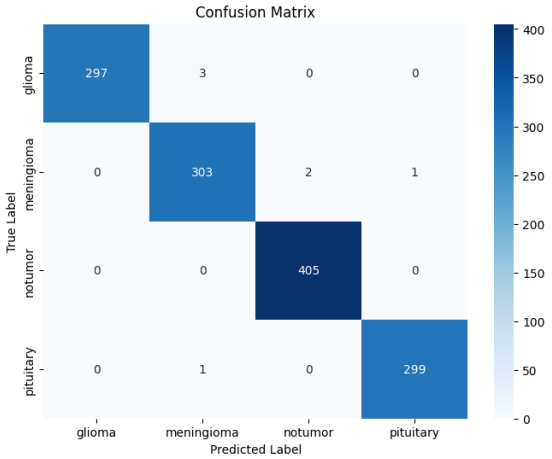

Fig. 3. Confusion matrix for the FAKD-XAI framework on the test set.

Comparison with State-of-the-Art (SOTA): We assess the efficacy of our proposed FAKD-XAI model against recent SOTA techniques. Table 3 contrasts our results with conventional CNNs and other knowledge distillation (KD) methods on comparable brain tumor MRI datasets.

Our baseline student model, MobileNetV3-Large, already achieves a competitive accuracy of 99.08%, outperforming larger models like ResNet152 [24]. This

Method	Dataset	Accuracy (%)
Brain Tumor Classification		
ResNet152 [24]	Brain Tumor MRI Dataset	98.50
PDCNN [25]	Brain Tumor MRI Dataset	98.12
LCDEiT [26]	Figshare MRI Dataset	98.11
MobileNetV3 Large (Ours)	**Brain Tumor MRI Dataset**	**99.08**
Classification with Knowledge Distillation		
FedBrain-Distill [14]	Figshare MRI Dataset	94.38
KD (CNN-ViT) [27]	Brain Tumor MRI Dataset	97.00
DenseNet20+ResNet152V2 [28]	Brain Tumor MRI Dataset	98.01
FAKD-XAI (Ours)	**Brain Tumor MRI Dataset**	**99.47**

Table 3. Comparison of Brain Tumor Classification Models and their Accuracy

high baseline is attributed to the powerful pre-trained features of MobileNetV3 and the relatively clean, well-defined nature of the Brain Tumor MRI dataset used.

The primary reason for FAKD-XAI's exceptional 99.47% accuracy lies in our **hybrid distillation strategy**. Unlike traditional KD methods that only match output logits (e.g., [27] achieving 97.00%), FAKD-XAI incorporates an **intermediate feature-level distillation**. By forcing the student model's intermediate representations to mimic those of the powerful ResNet-50 teacher, we transfer rich, hierarchical feature knowledge that is crucial for distinguishing subtle pathological patterns in medical images. Our feature alignment module, with its 1x1 convolution, effectively bridges the architectural gap between the teacher and student, enabling this deeper knowledge transfer. This is a significant advantage over simpler KD approaches. As seen in Table 3, our method surpasses other recent distillation techniques, including the DenseNet-ResNet combination [28] which reached 98.01%, demonstrating the superiority of our targeted feature-aligned approach for this task.

Qualitative Analysis using LIME: We applied LIME (Local Interpretable Model-agnostic Explanations) to explain the decision-making process of our FAKD-XAI model. By use of a simplified model on input picture perturbations, LIME finds superpixels either positively or negatively influencing forecasts. Displayed in every visualisation are four panels: the original labelled MRI, the top positive contributors (green), the top negative contributors (red), and a combined overlay. Green areas in Figure 4 (Pituitary) correctly depict the center-lower brain location of the tumour. Green sections in Figure 5 (Notumor) emphasise normal brain architecture; red areas draw attention to perhaps unclear parts. Figure 6 (Meningioma) shows green superpixels marking the position of the peripheral tumour along the head. Green regions in Figure 7 (Glioma) help to highlight the particular form of the tumour inside the brain parenchyma. By

matching actual tumour sites or normal tissue characteristics, our model's focus on clinically relevant variables increases prediction accuracy.

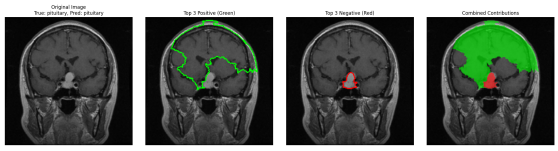

Fig. 4. LIME explanation for Pituitary tumor

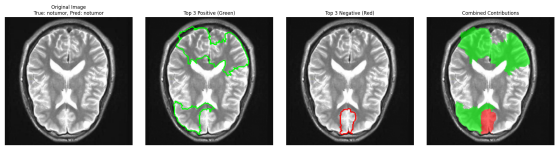

Fig. 5. LIME explanation for Notumor

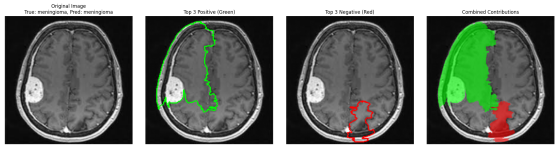

Fig. 6. LIME explanation for Meningioma

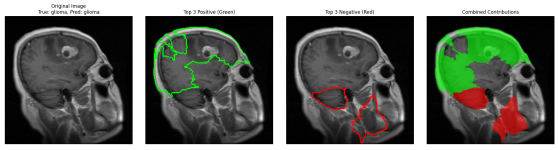

Fig. 7. LIME explanation for Glioma

4.4 Discussion

Our experimental findings show how well the FAKD-XAI framework performs classifying multi-class brain tumors from MRI scans. With 99.47% accuracy on the test dataset, the framework outperformed a number of contemporary SOTA techniques in Table 3. The lightweight MobileNetV3 Large student receives improved feature representation qualities by transferring information from the larger ResNet50 teacher model. Our hybrid method uses logit-level KD (by KL divergence) and feature-level KD (using MSE on projected, aligned feature maps). While feature KD compels comparable intermediate representations, hence producing a stronger student model, logit KD motivates the student to imitate the output probabilities of the teacher.

Qualitative study using LIME offers crucial understanding of the decision-making processes of models. Figures 4-7 illustrate the model's consistent identification of salient image regions associated with specific tumor types or characteristics of healthy tissue. The observed excellent performance metrics and the confidence in the forecasts are supported by the relationship between model attention and clinically important variables.

In similar tasks, FAKD-XAI shows better performance than traditional CNN methods and modern KD approaches, highlighting the effectiveness of the hybrid distillation approach for medical imaging uses. MobileNetV3 Large's achievement of state-of-the-art outcomes is remarkable given its efficient design, which enables possible use in resource-limited clinical environments.

Though promising, our study has limitations. Performance was assessed on a single dataset, warranting further validation. Additionally, our use of LIME, while providing intuitive local explanations, is a recognized limitation due to its potential instability. Future work should incorporate a comparative analysis with other XAI methods, such as Grad-CAM for visualizing feature importance and SHAP for more theoretically grounded explanations, to provide a more robust and comprehensive understanding of model behavior. The study also did not address potential dataset biases.

Future works include validation across several multi-institutional datasets, investigation of alternative architecture combinations, improvement of knowledge distillation loss components, research of complementary explainable AI techniques, and execution of clinical validation studies to evaluate real-world applications.

5 Conclusion

In this work, we presented **FAKD-XAI**, a novel knowledge distillation framework to improve the classification of brain tumors from MRI scans. Using a computationally efficient MobileNetV3-Large student model, our method successfully transfers knowledge from a complex ResNet-50 teacher network to MobileNetV3-Large, a lighter model. The fundamental innovation is the fusion of an adaptive feature alignment mechanism with traditional logit-level distillation,

which captures and transfers rich intermediate representations essential for identifying subtle pathological patterns. Our proposed approach set a new standard in the field by attaining a classification accuracy of **99.47%** through rigorous testing on the Brain Tumor MRI dataset. This performance significantly outperforms that of the baseline student model and other modern methodologies. The inclusion of **LIME** provides useful visual explanations, thereby improving model transparency and building confidence, which is a crucial element for clinical adoption. The FAKD-XAI framework efficiently balances model efficiency (5.25 ms inference time per image) with high accuracy (99.47%), proving its potential as a real-time diagnostic solution in resource-constrained medical settings. Future research will seek to confirm the framework's durability across multi-institutional datasets and to investigate complementary XAI technologies, such as Grad-CAM and SHAP, to build a more holistic and reliable interpretability framework.

References

1. National Brain Tumor Society, "Brain Tumor Facts," 2022. [Online]. Available: `https://braintumor.org/brain-tumors/about-brain-tumors/brain-tumor-facts/`, last accessed 2025/04/20.
2. Whelan, H.T., Clanton, J.A., Wilson, R.E., Tulipan, N.B.: Comparison of CT and MRI brain tumor imaging using a canine glioma model. Pediatric Neurology **4**(5), 279–283 (1988)
3. Mohsen, H., El-Dahshan, E.-S.A., El-Horbaty, E.-S.M., Salem, A.-B.M.: Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal **3**(1), 68–71 (2018)
4. Rohini, V., Kumar, K.P.: ConvNet based detection and segmentation of brain tumor from MR images. In: 2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 1–4. IEEE, Bengaluru, India (2021)
5. İşın, A., Direkoğlu, C., Şah, M.: Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Computer Science **102**, 317–324 (2016)
6. Karimi, E., Yu, M.W., Maritan, S.M., Perus, L.J.M., Rezanejad, M., Sorin, M., Dankner, M., Fallah, P., Doré, S., Zuo, D., et al.: Single-cell spatial immune landscapes of primary and metastatic brain tumours. Nature **614**(7948), 555–563 (2023)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
8. Romero, A., Ballas, N., Ebrahimi Kahou, S., Chassang, A., Gatta, C., Bengio, Y.: FitNets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
9. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Las Vegas, USA (2016)
11. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V.: Searching for MobileNetV3. In: IEEE International Conference on Computer Vision (ICCV), pp. 1314–1324. IEEE, Seoul, South Korea (2019)

12. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM, San Francisco, USA (2016)
13. Jiang, Y., Zhao, X., Wu, Y., Chaddad, A.: A Knowledge Distillation-Based Approach to Enhance Transparency of Classifier Models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 17653–17661. AAAI Press, Philadelphia, USA (2025)
14. Gohari, R.J., Aliahmadipour, L., Valipour, E.: FedBrain-Distill: Communication-Efficient Federated Brain Tumor Classification Using Ensemble Knowledge Distillation on Non-IID Data. arXiv preprint arXiv:2409.05359 (2024)
15. Anantathanavit, R., Raswa, F.H., Thaipisutikul, T., Wang, J.-C.: Lightweight Brain Tumor Diagnosis via Knowledge Distillation. In: 2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pp. 1–6. IEEE, Bangkok, Thailand (2024)
16. Kanchanamala, P., Kuppusamy, V., Ganesan, G.: QDCNN-DMN: A hybrid deep learning approach for brain tumor classification using MRI images. Biomedical Signal Processing and Control **101**, 107199 (2025)
17. Zarenia, E., Far, A.A., Rezaee, K.: Automated multi-class MRI brain tumor classification and segmentation using deformable attention and saliency mapping. Scientific Reports **15**(1), 8114 (2025)
18. Guan, Y., Aamir, M., Rahman, Z., Ali, A., Abro, W.A., Dayo, Z.A., Bhutta, M.S., Hu, Z.: A framework for efficient brain tumor classification using MRI images. *Math. Biosci. Eng.* **18**(5), 5790–5815 (2021). https://doi.org/10.3934/mbe.2021292
19. Chaitanya, P.S., Satpathy, S.K.: Advancing Brain Tumour Detection and Classification: Knowledge Distilled ResNeXt Model for Multi-Class MRI Analysis. International Journal of Computational and Experimental Science and Engineering **10**(4), 1610–1623 (2024). https://doi.org/10.22399/ijcesen.730
20. Nickparvar, M.: Brain Tumor MRI Dataset. Kaggle (2021). https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, last accessed 2025/04/18. https://doi.org/10.34740/KAGGLE/DSV/2645886
21. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
24. Mathivanan, S.K., Sonaimuthu, S., Murugesan, S., Rajadurai, H., Shivahare, B.D., Shah, M.A.: Employing deep learning and transfer learning for accurate brain tumor detection. Scientific Reports **14**(1), 7232 (2024)
25. Rahman, T., Islam, M.S.: MRI brain tumor detection and classification using parallel deep convolutional neural networks. Measurement: Sensors **26**, 100694 (2023)
26. Ferdous, G.J., Sathi, K.A., Hossain, M.A., Hoque, M.M., Dewan, M.A.A.: LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. IEEE Access **11**, 20337–20350 (2023)
27. Tabassum, M.: Visual Interpretation of Brain Tumor Detection Using Knowledge Distillation. PhD thesis, National University of Sciences and Technology (2023)
28. Khan, S.U.R., Asim, M.N., Vollmer, S., Dengel, A.: Robust & Precise Knowledge Distillation-based Novel Context-Aware Predictor for Disease Detection in Brain and Gastrointestinal. arXiv preprint arXiv:2505.06381 (2025)