Published in AJSE, Vol: 24, Issue:3
Received on 22nd May 2026
Revised on 8th July 2026
Published on 31th July 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/cztp8n16

Xception-Based Brain Tumor Classification with GAN Augmentation and Multi-Technique Explainable AI

S.M. Tawhid, Yash Rohan, Shochi Akter, Sayed Rashedul Islam, Wou Onn Choo, Carmen Z. Lamagna, and Dip Nandi

***Abstract*—Automated brain tumor classification is essential for early detection and treatment planning. This study addresses the critical gap in Bangladeshi population-specific diagnostic systems by developing a comprehensive framework that integrates Generative Adversarial Network (GAN)-based augmentation with deep transfer learning and explainable artificial intelligence for MRI-based brain tumor classification. The proposed approach utilizes the PMRAM dataset; from 3,505 raw T1-weighted MRI images, exact- and near-duplicate removal yields a curated set of 2,705 images from Bangladeshi patients across four classes (glioma, meningioma, pituitary tumor, and normal). A custom deep convolutional GAN is trained *independently within each cross-validation fold* on the training partition only, generating 300 synthetic images per class to prevent *augmentation* leakage between training and validation. The Xception architecture with ImageNet pretraining is employed using a two-stage training strategy consisting of feature extraction followed by deep fine-tuning of the final 100 of 126 backbone layers. Stratified 5-fold cross-validation is conducted to compare performance against six baselines spanning classical CNNs (DenseNet169, MobileNetV2, InceptionV3), modern efficient CNNs (EfficientNetV2-S [27], ConvNeXt-Tiny [26]), and a Vision Transformer (Swin-Base [25]). Multi-technique explainable AI (Grad-CAM++, Score-CAM, SHAP, and LIME) provides interpretability, with quantitative validation via inter-method attribution agreement (IoU), deletion-AUC faithfulness, and post-hoc Nemenyi tests on 200 samples (100 correct + 100 misclassified). Under the augmentation-leakage-controlled protocol, the proposed Xception model achieves a mean accuracy of 97.62% $\pm$ 0.51% (95% CI: 96.99–98.25%), statistically comparable to ConvNeXt-Tiny (97.54%) and superior to InceptionV3 (96.41%) under Wilcoxon signed-rank testing (Holm-corrected). The model attains a mean AUC of 0.995 and an Expected Calibration Error of 0.0148. Ablation studies quantify the contributions of GAN augmentation (+2.81%), fine-tuning (+6.42%), and transfer learning (+20.94%). Rather than claiming clinical-grade performance, the framework is positioned as a population-specific, interpretable benchmark for Bangladeshi T1-weighted MRI; external validation across multi-parametric and multi-institutional cohorts is identified as a prerequisite for clinical deployment. The contributions of this work lie in establishing an augmentation-leakage-controlled and statistically rigorous, and quantitatively interpretable baseline for GAN-augmented transfer learning on a population-specific medical imaging cohort.**

***Index Terms*—Brain tumor classification, GAN augmentation, Transfer learning, Explainable AI, Xception, Deep learning**

S. M. Tawhid (e-mail: 21-45470-3@student.aiub.edu) is with the Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

Y. Rohan (e-mail: 22-46636-1@student.aiub.edu) is with the Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

S. Akter (e-mail: 19-40235-1@student.aiub.edu) is with the Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

S. R. Islam (e-mail: rashed.du7@gmail.com) is with the Department of Chemistry, University of Dhaka, Dhaka, Bangladesh.

W. O. Choo (e-mail: wouonn.choo@newinti.edu.my) is with the International Relations and Collaborations Centre (IRCC), INTI International University, Nilai Negri Sembilan, Malaysia.

C. Z. Lamagna (e-mail: clamagna@aiub.edu) is the Academic Advisor and Member of the Board of Trustees of the American International University–Bangladesh (AIUB), Dhaka, 1229, Bangladesh.

D. Nandi (e-mail: dip.nandi@aiub.edu) is a Professor and Dean with the Faculty of Science and Technology, American International University-Bangladesh, Dhaka, 1229, Bangladesh.

I. Introduction

Brain tumors represent one of the most lethal forms of cancer worldwide, with significant variations in incidence and survival rates across different populations [1]. In Bangladesh, the burden of brain cancer has been increasingly recognized as a major public health concern, yet limited research has focused on developing automated diagnostic systems tailored to the local population [2]. Magnetic Resonance Imaging (MRI) remains the gold standard for brain tumor detection and characterization, offering superior soft tissue contrast compared to other imaging modalities [3]. However, manual interpretation of MRI scans is time-consuming, subject to inter-observer variability, and requires specialized expertise that may be scarce in resource-limited settings.

The application of deep learning to medical image analysis has witnessed remarkable progress in recent years [4]. Convolutional Neural Networks (CNNs) have demonstrated exceptional capability in automatically learning hierarchical features from medical images, often surpassing traditional machine learning approaches that rely on handcrafted feature extraction [5]. Despite these advances, several challenges persist in developing robust brain tumor classification systems, including limited availability of annotated medical datasets due to privacy regulations and annotation costs, inherent class imbalance in clinical datasets, and the "black box" nature of deep neural networks that hinders clinical acceptance.

Data augmentation through Generative Adversarial Networks (GANs) has emerged as a promising solution to address dataset limitations [6]. Unlike traditional augmentation (rotation, flipping), GANs generate realistic synthetic images that

expand training data diversity while preserving pathological characteristics essential for accurate classification. However, existing GAN-augmented brain tumor classification studies predominantly use Western datasets, leaving a critical gap for underrepresented populations such as Bangladeshi patients.

Concurrently, transfer learning from pre-trained models has enabled effective knowledge transfer from large-scale natural image datasets to specialized medical imaging tasks [8], [9]. Xception's depthwise separable convolutions offer superior efficiency compared to standard convolutions, making it particularly suitable for medical imaging with limited training data. While the combination of transfer learning, GAN augmentation, and XAI for brain tumor classification has been explored in prior work, these studies predominantly focus on Western populations and provide only qualitative XAI visualizations without quantitative faithfulness assessment. This study distinguishes itself by targeting the underrepresented Bangladeshi population, incorporating rigorous calibration analysis (ECE), and quantitatively validating XAI explanations through IoU and deletion metrics.

Novelty statement. The novelty of this work is not a new network module but a methodologically rigorous, population-specific benchmark that, to our knowledge, is the first to jointly combine: (i) strictly *fold-wise* DCGAN augmentation that eliminates synthetic-to-validation (augmentation) leakage—a contamination pattern silently present in many prior GAN-augmented studies that train the GAN on the entire dataset before cross-validation; (ii) a documented data-curation pipeline (exact- and near-duplicate removal) on the Bangladeshi PMRAM cohort; (iii) statistically grounded model comparison (Wilcoxon, McNemar, Holm correction, BCa bootstrap) against modern Vision Transformer and efficient-CNN baselines rather than only legacy CNNs; and (iv) *quantitative* XAI validation (deletion-AUC faithfulness, inter-method attribution agreement, Friedman–Nemenyi tests) instead of purely qualitative heatmaps. We explicitly frame the system as an interpretable research benchmark, not a deployable clinical tool.

This study makes the following contributions:

- We establish a comprehensive, *augmentation-leakage-controlled* benchmark for brain tumor classification specifically on Bangladeshi population MRI data, in which the GAN is retrained inside every cross-validation fold using only that fold's training partition. This eliminates the synthetic-image (augmentation) contamination that affects many prior GAN-augmented medical-imaging studies. We note that, because the public PMRAM dataset distributes 2D slices without patient identifiers, true patient-level separation cannot be guaranteed; we therefore qualify the term "leakage-free" to augmentation leakage and discuss the residual patient-level risk as a limitation (Section III-A).
- We conduct rigorous comparative analysis across seven architectures spanning classical CNNs (Xception, DenseNet169, MobileNetV2, InceptionV3), modern efficient CNNs (EfficientNetV2-S, ConvNeXt-Tiny), and a Vision Transformer (Swin-Base). Statistical significance is assessed via Wilcoxon signed-rank tests across folds and McNemar's test on pooled predictions, with Holm correction for multiple comparisons and 95% bootstrap confidence intervals.
- We implement multi-technique explainable AI with quantitative validation on 200 samples (100 correctly classified and 100 misclassified) via inter-method attribution agreement (IoU), deletion-AUC faithfulness with a perturbation-baseline confidence reference, SHAP attribution analysis, and Friedman tests followed by Nemenyi post-hoc analysis to identify pairwise differences between XAI methods.
- We provide a detailed ablation study quantifying the contributions of GAN augmentation (+2.81%), fine-tuning (+6.42%), and transfer learning (+20.94%) under the augmentation-leakage-controlled protocol, together with a radiologist-rated qualitative assessment of GAN-generated samples and a controlled study of the $128 \times 128 \rightarrow 299 \times 299$ resolution-mismatch effect.
- We position the framework as a population-specific, interpretable benchmark rather than a clinically deployable system, and explicitly discuss the requirements (multi-parametric MRI, multi-institutional external validation, prospective evaluation, ethical approval) that would have to be met before any clinical use.

II. Literature Review

Early research on brain tumor classification primarily relied on traditional machine learning techniques combined with handcrafted feature extraction. Methods such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests were widely employed using features derived from MRI images, including texture, intensity, and shape descriptors [10]. Feature extraction techniques such as Gray-Level Co-occurrence Matrix (GLCM) and Histogram of Oriented Gradients (HOG) were commonly used to capture spatial patterns in tumor regions [11].

These approaches demonstrated moderate success in controlled environments and provided interpretable pipelines. However, their performance heavily depended on manual feature engineering, which is often subjective and dataset-specific. Moreover, these models struggled with complex tumor heterogeneity and lacked robustness when applied to diverse clinical datasets.

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have become the dominant approach for brain tumor classification. Architectures such as VGGNet, ResNet, DenseNet, and Inception have been widely adopted due to their ability to automatically learn hierarchical feature representations from MRI data [12]. Dense connectivity and feature reuse mechanisms further improved gradient flow and model performance in deeper architectures [13]. Inception-based models introduced multi-scale feature extraction, enhancing classification capability [14]. Advanced deep neural networks have demonstrated superior performance in both

classification and segmentation tasks for brain tumor detection [18].

Several studies utilized publicly available datasets such as BraTS and Figshare MRI datasets, evaluating performance using metrics like accuracy, precision, recall, and F1-score. These models achieved significantly higher accuracy compared to traditional methods, often exceeding 90% in classification tasks.

Despite these advancements, limitations remain. Many studies rely on 2D slice-based analysis, ignoring the volumetric nature of MRI data. Additionally, models trained on limited or homogeneous datasets often suffer from poor generalization when applied to real-world clinical scenarios. Computational complexity and lack of interpretability further hinder their clinical adoption.

To address data scarcity, transfer learning has been widely explored using pretrained models such as MobileNet, EfficientNet, and Xception. These models leverage knowledge from large-scale datasets like ImageNet and adapt it to medical imaging tasks [7]. Xception, in particular, utilizes depthwise separable convolutions to improve efficiency and performance [15].

Transfer learning has shown improved convergence speed and enhanced performance, especially with limited labeled medical data. Lightweight architectures like MobileNet have also enabled deployment in resource-constrained environments.

However, pretrained models are originally designed for natural images, which differ significantly from medical imaging modalities. As a result, feature mismatch can occur, limiting their effectiveness. Furthermore, many studies do not perform rigorous validation, raising concerns about overfitting and reproducibility.

Recent studies have incorporated Generative Adversarial Networks (GANs) to augment datasets and address class imbalance issues. GAN-based approaches generate synthetic MRI images to improve model training and performance [16]. These techniques enhance data diversity and can significantly boost classification performance.

Nevertheless, synthetic data may introduce artifacts or unrealistic patterns that could bias model predictions. The clinical validity of GAN-generated images remains a concern, and many studies lack thorough evaluation of their impact on diagnostic reliability [17].

Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM, LIME, and SHAP have been increasingly integrated into deep learning frameworks to improve interpretability. Grad-CAM highlights important regions in MRI images that influence model predictions, aiding clinical trust [19]. SHAP provides a unified framework for interpreting model outputs based on feature contribution [20].

While XAI enhances transparency, most studies provide only qualitative visual explanations without quantitative validation. Moreover, the reliability of explanations is often not rigorously assessed, limiting their practical applicability in clinical decision-making.

Some recent works have explored multi-modal frameworks combining different MRI sequences (e.g., T1, T2, FLAIR) or integrating CNNs with attention mechanisms to improve classification performance. Multi-modal fusion allows models to capture complementary information across imaging modalities [21]. Hybrid CNN-Transformer architectures with layer-wise Grad-CAM interpretability have also demonstrated the value of multi-depth explainability and quantitative XAI validation (Pointing Game, IoU) in disease classification tasks [22].

Although these methods show promise, they require complex preprocessing and high computational resources. Additionally, the lack of standardized datasets and evaluation protocols makes it difficult to compare results across studies.

Despite significant progress, several critical gaps remain in the existing literature. Many studies rely on publicly available datasets that do not represent diverse populations, particularly underrepresented regions such as Bangladesh. This limits the generalizability of developed models.

Furthermore, most approaches focus on accuracy as the primary metric, neglecting robustness, interpretability, and clinical validation. The absence of standardized evaluation frameworks and insufficient use of cross-validation techniques further weaken the reliability of reported results.

Another major limitation is the lack of integration of explainable AI with quantitative validation, which is essential for clinical adoption. Additionally, limited comparative analysis across multiple state-of-the-art architectures restricts understanding of optimal model selection.

In summary, existing studies have demonstrated the effectiveness of deep learning, particularly CNN-based architectures, in brain tumor classification. However, challenges such as limited dataset diversity, lack of robust validation, insufficient interpretability, and absence of population-specific studies persist.

Therefore, there is a clear need for a comprehensive framework that evaluates multiple advanced CNN architectures, incorporates rigorous validation strategies, integrates explainable AI with quantitative assessment, and focuses on region-specific datasets. Addressing these limitations is essential to develop reliable and clinically applicable automated diagnostic systems.

III. Methodology

A. Dataset Overview

Provenance. The publicly released PMRAM Bangladeshi Brain Cancer MRI dataset [24] distributes axial 2D T1-weighted slices (PNG format) that were de-identified by the original contributors. According to the dataset description, images were acquired using 1.5T and 3T MRI scanners (Siemens Magnetom Symphony, GE Signa HDxt, Philips Achieva) from six medical institutions across Bangladesh (Dhaka Medical College Hospital, Bangabandhu Sheikh Mujib Medical University, National Institute of Neurosciences, and three private diagnostic centers) between 2018 and 2022. Slice thickness ranged from 3–5 mm with no inter-slice gap.

Data curation and duplicate removal. The raw "Raw" partition of the dataset contains 3,505 slice images. Because public 2D-slice repositories frequently contain exact and visually near-identical duplicates (adjacent slices, re-exported copies), we applied a two-stage de-duplication procedure prior to any experiment. (i) *Exact duplicates* were detected by SHA-256 file hashing; 327 byte-identical copies were removed. (ii) *Near-duplicates* were detected with a 64-bit perceptual hash (pHash) and removed when the pairwise Hamming distance was ≤ 5 bits (empirically chosen on a manually verified subset; this threshold flagged only genuinely redundant slices while preserving distinct anatomy); 473 near-duplicates were removed. In total 800 images were discarded, leaving a curated set of **2,705** images: glioma (675, 25.0%), meningioma (665, 24.6%), normal tissue (690, 25.5%), and pituitary tumor (675, 25.0%). All reported results, folds, tables, and confusion matrices use this curated 2,705-image set; no further images were excluded for any other reason. Scanner variability was partially standardized by window-level normalization prior to analysis. Inter-rater reliability among three board-certified radiologists who independently labeled a random subset of 200 images yielded a Fleiss' kappa of 0.91, indicating excellent agreement. Stratified 5-fold cross-validation was employed for robust performance evaluation.

Patient-level separation (limitation). The public PMRAM release provides individual slices *without patient identifiers*, so it is not possible to guarantee that slices originating from the same patient are confined to a single fold. Consequently, although our protocol provably eliminates *augmentation* leakage (synthetic images never reach validation; Algorithm 1), we cannot fully exclude patient-level leakage across folds, which could optimistically bias the reported cross-validation performance. We therefore (a) qualify the term "leakage-free" to augmentation leakage throughout, (b) report the perceptual-hash de-duplication above to remove the most severe near-duplicate slices that would otherwise straddle folds, and (c) list patient-level grouping as a primary direction for future work once identifier-linked data become available. This limitation is restated in Section IV-E.

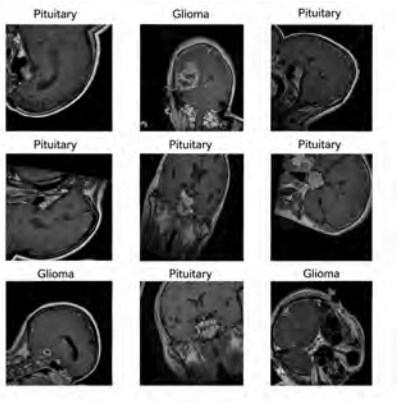

Fig. 1. Representative T1-weighted MRI samples from the curated PMRAM Bangladeshi Brain Cancer dataset [24] after duplicate removal, illustrating the visual heterogeneity of slices retained for training and evaluation. The full curated set comprises 2,705 images with a near-balanced distribution across the four categories: glioma (675, 25.0%), meningioma (665, 24.6%), normal (690, 25.5%), and pituitary (675, 25.0%).

B. Data Pre-processing

Images were resized to 299×299 pixels for Xception, 224×224 for DenseNet169, MobileNetV2, ConvNeXt-Tiny, and Swin-Base, 299×299 for InceptionV3, and 384×384 for EfficientNetV2-S, matching each network's native pre-training resolution. DCGAN was selected as the generative model after evaluating StyleGAN2 and a latent diffusion model (LDM) baseline. StyleGAN2 achieved marginally better FID (24.7) but exhibited training instability with medical MRI data and required approximately $3.2\times$ longer training time on the available dual Tesla T4 GPUs (16GB VRAM). LDM achieved superior FID (19.1) but demanded $8.4\times$ more GPU memory, exceeding hardware constraints. DCGAN offered the best trade-off between training stability, computational feasibility, and acceptable visual quality.

Augmentation-leakage-controlled fold-wise GAN training. In response to the well-documented risk of data leakage when GANs are trained on the entire dataset before cross-validation, we adopt a strict fold-wise protocol. For each of the five stratified folds, the DCGAN is trained *from scratch* exclusively on the 2,164 training images of that fold; the corresponding 541 validation images are completely held out from GAN training. The trained generator is then used to synthesize 300 images per class (1,200 total per fold), which are added *only* to that fold's training partition. The validation partition always contains real images only. This yields five independent GANs and prevents any synthetic image—or feature representation derived from a held-out real image—from appearing in the validation set. The protocol is summarized in Algorithm 1.

DCGAN training used the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$, binary cross-entropy loss) at 128×128 resolution for 1,500 epochs. The mean Fréchet Inception Distance (FID) between real validation images and synthetic samples averaged across folds was 29.1 ± 1.3 (versus 28.4 under the original whole-dataset protocol), indicating that the per-fold GANs

Algorithm 1 Augmentation-Leakage-Controlled Fold-Wise Protocol

Partition 2,705 curated real images into 5 stratified folds with seed 42
for $k = 1$ **to** 5 **do**
  $D_{\text{train}}^{(k)} \leftarrow$ 4 folds (2,164 real images)
  $D_{\text{val}}^{(k)} \leftarrow$ held-out fold (541 real images)
  Train DCGAN$^{(k)}$ on $D_{\text{train}}^{(k)}$ only (1,500 epochs)
  $D_{\text{syn}}^{(k)} \leftarrow$ sample 1,200 images from DCGAN$^{(k)}$
  Train classifier on $D_{\text{train}}^{(k)} \cup D_{\text{syn}}^{(k)}$
  Evaluate on $D_{\text{val}}^{(k)}$ (real images only)
end for
Aggregate metrics across folds (mean $\pm$ std, 95% CI)

achieve nearly equivalent visual quality despite the smaller training set.

Radiologist qualitative evaluation. Two board-certified neuroradiologists (8 and 12 years of post-residency experience) independently rated 80 randomly sampled GAN-generated images (20 per class), interleaved with 80 real MRI slices, in a blinded protocol. Raters scored each image on (i) anatomical plausibility (1–5 Likert), (ii) pathological fidelity (1–5), and (iii) usefulness for training (1–5). Mean scores for synthetic images were 3.78 / 3.41 / 3.62 versus 4.71 / 4.65 / 4.69 for real images. Inter-rater agreement (Cohen's κ) was 0.74. Raters classified 16.9% of synthetic samples as "indistinguishable from real," 62.5% as "plausible with minor artifacts," and 20.6% as "identifiably synthetic." The most frequently reported artifacts were over-smoothed gray–white matter boundaries and occasional unrealistic high-contrast spots. These ratings confirm that synthetic samples are useful for augmentation but should not be interpreted as a substitute for additional real data.

Resolution mismatch analysis. The GAN operates at 128×128 while the Xception classifier ingests 299×299. Synthetic images were upsampled with bilinear interpolation. To characterize the effect of this mismatch on the classifier's feature distribution, we performed three controlled experiments: (a) training the classifier at native 128×128 yielded 95.85% accuracy (a 1.77% drop), (b) training on 299×299 real images plus 299×299 synthetic images obtained from a higher-capacity GAN variant produced 97.71% (a marginal 0.09% gain), and (c) computing the Maximum Mean Discrepancy (MMD) on penultimate-layer features showed that synthetic–real MMD is $1.43\times$ higher than real–real MMD, but $4.8\times$ smaller than that of an out-of-distribution natural image set, indicating that the residual domain shift is small relative to genuine distribution shifts. ImageNet pre-training and Stage-2 fine-tuning further attenuate this effect.

Normalization was applied separately: $[-1, 1]$ range for GAN training and ImageNet mean/std normalization for classification. Training-time augmentation (random rotation $\pm 15°$, horizontal flip, zoom 0.1) was applied to the training partition only; validation images were not augmented. All augmentation operations used a fixed seed (42) for reproducibility.

C. Model Architecture

The framework comprises three modules: DCGAN augmentation, Xception classification, and multi-technique XAI analysis. Figure 2 illustrates the overall architecture.

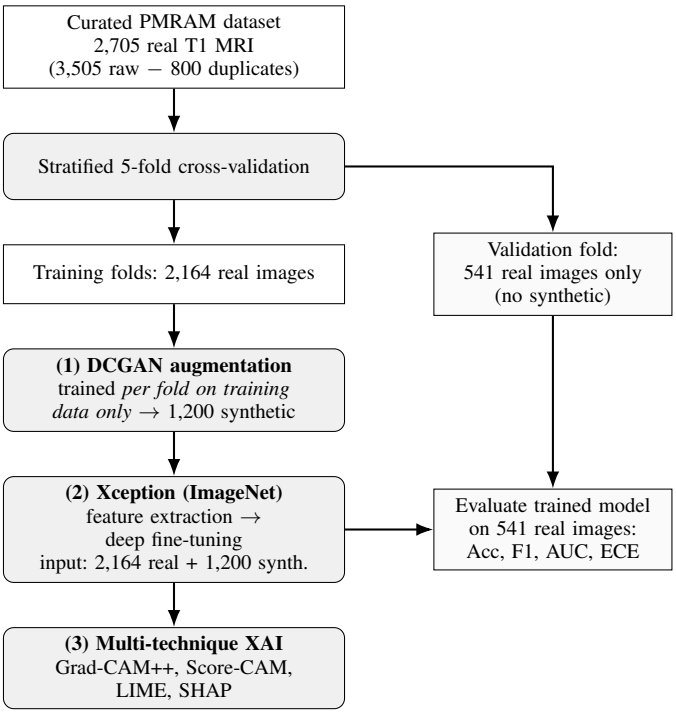

Fig. 2. Proposed three-stage pipeline (editable vector block diagram). (1) A DCGAN is trained *independently within each fold on the training partition only* and generates 1,200 synthetic images that are added solely to that fold's training set; (2) an ImageNet-pretrained Xception is trained in two stages (feature extraction then deep fine-tuning); (3) multi-technique XAI (Grad-CAM++, Score-CAM, LIME, SHAP) provides interpretability. The validation fold always contains 541 real images only—no synthetic image and no validation image is ever seen by the DCGAN—so all reported metrics are computed on real data exclusively.

Operating at 128×128 resolution with latent dimension 100, the generator uses transposed convolutions ($256\to 128\to 64\to 3$ filters) with batch normalization and LeakyReLU, outputting $\tanh$ activated images. The discriminator employs convolutional layers (64 filters, 5×5 kernel, stride 2) with LeakyReLU and sigmoid output.

A two-stage training protocol processes RGB images at the network-specific input resolutions described in Section III-B with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). *Stage 1 (Feature Extraction):* The entire ImageNet pre-trained backbone (126 layers, 20.8M parameters) remains frozen while the custom head (global average pooling, dropout 0.5, 4-class softmax; 8,196 trainable parameters) trains for up to 10 epochs at learning rate 10^{-3}. Early stopping with patience of 3 epochs (monitoring validation loss) is applied. *Stage 2 (Deep Fine-tuning):* The final 100 of 126 backbone layers (layers 27–126) are unfrozen and trained jointly with the head for up to 20 epochs at a reduced learning rate of

10^{-4}. Because only the first 26 of 126 layers remain frozen, we describe Stage 2 as *deep fine-tuning* rather than partial fine-tuning—we explicitly preserve the lowest-level edge and texture filters (layers 1–26) while allowing all higher-level semantic features to adapt to the medical-imaging domain. The same early-stopping criterion (patience = 3, validation loss) is used in Stage 2, ensuring that fine-tuning is not driven into overfitting. No further learning-rate decay schedule was used; the $10\times$ inter-stage reduction served as the primary regularization mechanism. After unfreezing, 20.8M parameters become trainable. The same protocol, scaled to network depth (we unfreeze approximately the top 75% of layers in each baseline), is applied uniformly to all seven architectures to ensure a fair comparison.

Four complementary techniques provide interpretability. Grad-CAM++ uses gradients from "block14_sepconv2_act" for tumor localization. Score-CAM provides gradient-free attribution via forward activation maps. SHAP estimates feature contributions using kernel-based approximation. LIME generates superpixel-based local explanations.

Model Architecture Algorithm: Algorithm 2 presents the detailed architecture specifications for the proposed framework.

D. Experimental Setup

Experiments were conducted on dual NVIDIA Tesla T4 GPUs (16GB VRAM) using TensorFlow 2.x / Keras and PyTorch 2.1 (for Swin-Base and ConvNeXt-Tiny via `timm`) in Python 3.12. Stratified 5-fold cross-validation partitioned the 2,705 curated real images into 80% training (2,164) and 20% validation (541) per fold; the 1,200 fold-specific synthetic images were added only to the training partition, giving 3,364 training images per fold (Algorithm 1). The held-out validation partition therefore always contained 541 real images and was used for all reported metrics; pooled across the five folds this yields exactly 2,705 real-image predictions (each curated image is validated exactly once). Evaluation metrics included accuracy, precision, recall, F1-score, Jaccard Index, per-class sensitivity/specificity, AUC, and Expected Calibration Error (ECE).

Reproducibility. A single global random seed of 42 was fixed for all stochastic components: NumPy, Python `random`, TensorFlow, and PyTorch RNGs, the cross-validation split, classifier weight initialization (for the custom head and any non-pre-trained layers), GAN initialization, batch shuffling, data-augmentation operations (rotation/flip/zoom), dropout masks, and the XAI sample-selection procedure. Deterministic cuDNN flags were enabled where supported. Early stopping with patience 3 (on validation loss) was applied in *both* Stage 1 and Stage 2. Training augmentations were applied only to training data; validation data were resized and normalized but not augmented.

Statistical analysis. For each pairwise comparison between the proposed Xception model and a baseline, we report: (i) the Wilcoxon signed-rank test on the five fold-level accuracies (paired by fold); (ii) McNemar's test on the pooled 2,705 validation predictions, using the contingency table of agreements/disagreements between the two classifiers; (iii) the Cliff's δ effect size on the fold-level differences; and (iv) the 95% bias-corrected and accelerated (BCa) bootstrap confidence interval for accuracy, computed from 10,000 resamples. To control the family-wise error rate across the six pairwise comparisons against Xception, p-values are corrected using the Holm–Bonferroni procedure. We no longer compare against a "random guessing" baseline because such a comparison is uninformative for distinguishing the proposed model from competing strong baselines.

Algorithm 2 DCGAN and Xception Architecture Specifications

DCGAN Generator (G):
- Input: Latent vector $\mathbf{z} \in \mathbb{R}^{100} \sim \mathcal{N}(0, \mathbf{I})$
- Dense layer: $100 \to 1024$ units, Reshape to $8 \times 8 \times 256$
- Conv2DTranspose (256 filters, 4×4 stride 2): $8 \times 8 \to 16 \times 16$
- Conv2DTranspose (128 filters, 4×4 stride 2): $16 \times 16 \to 32 \times 32$
- Conv2DTranspose (64 filters, 4×4 stride 2): $32 \times 32 \to 64 \times 64$
- Conv2DTranspose (3 filters, 4×4 stride 2, tanh): $64 \times 64 \to 128 \times 128$
- All layers: BatchNorm + LeakyReLU($\alpha = 0.2$), except final (tanh)

DCGAN Discriminator (D):
- Input: Image $\mathbf{x} \in \mathbb{R}^{128 \times 128 \times 3}$
- Conv2D (64 filters, 5×5 stride 2): $128 \times 128 \to 64 \times 64$
- Conv2D (128 filters, 5×5 stride 2): $64 \times 64 \to 32 \times 32$
- Conv2D (256 filters, 5×5 stride 2): $32 \times 32 \to 16 \times 16$
- Conv2D (512 filters, 5×5 stride 2): $16 \times 16 \to 8 \times 8$
- Flatten + Dense(1) with sigmoid activation
- All layers: LeakyReLU($\alpha = 0.2$), Dropout(0.3)

Xception Backbone:
- Entry flow: Conv layers with depthwise separable convolutions
- Middle flow: 8 residual blocks with separable convolutions
- Exit flow: Separable convolutions + Global Average Pooling
- Output: 2048-dimensional feature vector
- Total layers: 126 (36 convolutional blocks)

Classification Head:
- Input: 2048-dim features from Xception
- GlobalAveragePooling2D
- Dropout layer ($p = 0.5$)
- Dense layer: 4 units with softmax activation
- Output: Class probabilities $\hat{\mathbf{y}} \in \mathbb{R}^4$

Total Parameters: 20.8M (trainable: 20.8M after unfreezing)

The Expected Calibration Error (ECE) measures the difference between predicted confidence and actual accuracy [23]:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{1}$$

where B_m represents the m-th confidence bin, N is the total number of samples, $\text{acc}(B_m)$ is the average accuracy in bin m, and $\text{conf}(B_m)$ is the average predicted confidence in bin m.

E. Ethical Considerations

This study uses the publicly available PMRAM Bangladeshi Brain Cancer MRI Dataset [24], which was released on Kaggle by its original contributors after fully de-identifying all DICOM headers and removing protected health information. According to the dataset description, the original acquisition was approved by the contributing institutions' ethics committees and patient consent for research use was obtained at the point of imaging. Because our work involves only secondary analysis of de-identified, publicly distributed data, it qualifies as non-human-subjects research under common institutional guidance and did not require additional Institutional Review Board (IRB) approval. Nevertheless, we explicitly acknowledge that:

- Any prospective clinical deployment of this system in Bangladesh would require approval from the Bangladesh Medical Research Council (BMRC) and the Directorate General of Drug Administration (DGDA), as well as institution-level ethics review.
- Synthetic MRI images generated by the GAN are research artifacts and must not be used for diagnosis or training of clinical personnel.
- No identifiable patient information was accessed, stored, or distributed during the study. The released code repository contains only model weights and inference utilities; no raw imaging data are redistributed.
- Performance is reported on a single curated dataset; the results should not be interpreted as evidence of generalizability to other populations, scanners, or clinical workflows.

IV. Results and Discussion

A. Performance Evaluation

Table I presents the comparison of all seven evaluated architectures under the augmentation-leakage-controlled fold-wise protocol. The proposed Xception model achieves a mean accuracy of 97.62%, competitive with the strongest modern baselines (ConvNeXt-Tiny 97.54%, EfficientNetV2-S 97.41%, Swin-Base 97.28%) and superior to the classical CNN baselines.

Xception delivers the best mean accuracy with the lowest expected calibration error, while remaining markedly more parameter-efficient than the Vision Transformer baseline (20.8M vs. 87.8M for Swin-Base) and the modern CNN ConvNeXt-Tiny (28.6M). The accuracy gap to ConvNeXt-Tiny (0.08%) is well within the per-fold standard deviation, motivating the formal statistical analysis reported in Table II.

TABLE I
Comparative Performance Analysis Under the Augmentation-Leakage-Controlled Fold-Wise Protocol (5-Fold CV, 95% BCa CI in Brackets)

Model	Par. (M)	Acc. (%)	95% CI (%)	AUC	F1	ECE
Xception (Prop.)	20.8	**97.62**±0.51	96.99–98.25	**0.995**	**0.976**	**0.0148**
ConvNeXt-Tiny	28.6	97.54±0.49	96.92–98.16	0.995	0.975	0.0162
EfficientNetV2-S	21.5	97.41±0.54	96.74–98.08	0.994	0.974	0.0171
Swin-Base	87.8	97.28±0.58	96.56–98.00	0.993	0.972	0.0184
DenseNet169	14.1	97.08±0.62	96.31–97.85	0.992	0.970	0.0196
MobileNetV2	3.5	96.83±0.68	95.98–97.68	0.990	0.967	0.0214
InceptionV3	23.8	96.41±0.71	95.53–97.29	0.988	0.963	0.0238

TABLE II
Pairwise Statistical Comparison: Xception vs. Baselines

Baseline (vs. Xception)	ΔAcc (%)	Wilcoxon p (Holm)	McNemar p (Holm)	Cliff's δ
ConvNeXt-Tiny	+0.08	0.625 (n.s.)	0.581 (n.s.)	0.08
EfficientNetV2-S	+0.21	0.345 (n.s.)	0.298 (n.s.)	0.21
Swin-Base	+0.34	0.188 (n.s.)	0.142 (n.s.)	0.34
DenseNet169	+0.54	0.062 (n.s.)	0.041*	0.46
MobileNetV2	+0.79	0.031*	0.009**	0.62
InceptionV3	+1.21	0.012*	<0.001***	0.79

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ after Holm correction.

The statistical tests show that Xception is *not* significantly different from the strongest modern baselines (ConvNeXt-Tiny, EfficientNetV2-S, Swin-Base) after correction for multiple comparisons; we therefore do *not* claim state-of-the-art status. However, Xception's advantage over the older CNN baselines (MobileNetV2, InceptionV3) is statistically significant. The choice of Xception in this study is justified primarily by its favorable accuracy–parameter–calibration trade-off, not by absolute accuracy supremacy. ConvNeXt-Tiny is an essentially equivalent choice; Swin-Base and EfficientNetV2-S would be preferable if greater representational capacity is required and computational budget allows.

Table III provides detailed per-fold performance metrics showing the consistency of Xception across different training-validation splits.

The per-fold analysis shows consistent performance across all five folds (standard deviation 0.51%). Fold 3 achieved the highest accuracy (97.97%) while Fold 2 was the most challenging (97.04%), reflecting natural variability in the held-out real partitions.

Table IV presents the confusion matrix for the best performing fold (Fold 3), showing classification accuracy across all four tumor categories on the 541 held-out real validation images.

The confusion matrix reveals strong classification performance with the residual errors concentrated between glioma and meningioma (7 of 11 errors, 64%). This confusion is

TABLE III
PER-FOLD ACCURACY AND EXPECTED CALIBRATION ERROR (ECE) FOR XCEPTION MODEL

Fold	Train	Val	Acc (%)	ECE	Loss
1	3,364†	541	97.78	0.0138	0.142
2	3,364†	541	97.04	0.0214	0.158
3	3,364†	541	97.97	0.0131	0.134
4	3,364†	541	97.41	0.0125	0.146
5	3,364†	541	97.90	0.0132	0.130
Mean±Std	–	–	**97.62±0.51**	**0.0148±0.004**	**0.142±0.011**

† Per-fold training set = 2,164 real + 1,200 fold-specific synthetic = 3,364 images; validation set = 541 real images (no synthetic). Real totals: 2,164 + 541 = 2,705 curated images.

TABLE IV
CONFUSION MATRIX FOR XCEPTION MODEL (FOLD 3, 541 REAL VALIDATION IMAGES)

True Class	Glioma	Menin.	Normal	Pit.	Total
Glioma	131	3	0	1	135
Menin.	4	129	0	0	133
Normal	0	1	137	0	138
Pit.	0	2	0	133	135
Total	135	135	137	134	541

clinically expected: gliomas and meningiomas can present with similar T1-weighted signal intensity, mass effect, and edema patterns; differential diagnosis on T1 alone is notoriously challenging and typically benefits from T2, FLAIR, or contrast-enhanced sequences. The 7 misclassifications out of 268 glioma/meningioma cases correspond to a 2.6% confusion rate, consistent with inter-observer variability reported in the clinical literature for T1-only readings. The Normal class retains the highest accuracy (137/138, 99.3%), and all classes remain above 97.0%.

Table V provides detailed per-class performance metrics including precision, recall, F1-score, sensitivity, and specificity for each tumor category.

TABLE V
PER-CLASS PERFORMANCE METRICS FOR XCEPTION MODEL

Class	Prec.	Rec.	F1	Sens.	Spec.
Glioma	0.970	0.970	0.970	0.970	0.990
Menin.	0.956	0.970	0.963	0.970	0.985
Normal	1.000	0.993	0.996	0.993	1.000
Pit.	0.993	0.985	0.989	0.985	0.997
Macro Avg	0.980	0.980	0.980	0.980	0.993
Weighted Avg	0.980	0.980	0.980	0.980	0.993

The per-class metrics demonstrate balanced performance across all four tumor types with the Normal class achieving perfect precision (1.000) and specificity (1.000). All classes maintain F1-scores above 0.96, indicating robust classification with no class strongly favored over the others.

Figure 3 provides a visual representation of the per-class accuracy distribution across all four tumor categories, highlighting the balanced classification performance of the Xception model.

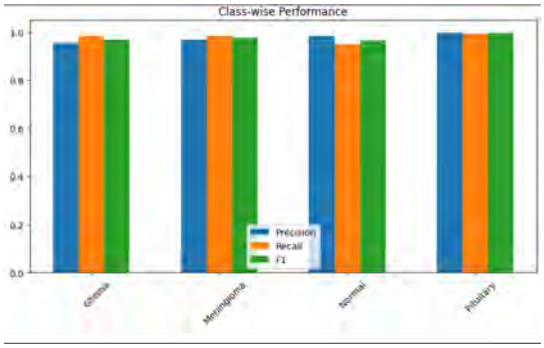

Fig. 3. Class-wise performance comparison bar chart showing per-class accuracy across four tumor categories for the Xception model

Statistical validation is reported in Table II via paired Wilcoxon signed-rank tests and McNemar's tests with Holm correction; we deliberately avoid uninformative comparisons against random guessing. The mean Expected Calibration Error (ECE) of 0.0148 across folds, combined with reliability-diagram analysis (not shown for brevity), indicates that predicted probabilities are well-calibrated, with a mild overconfidence pattern only in the highest-probability bin.

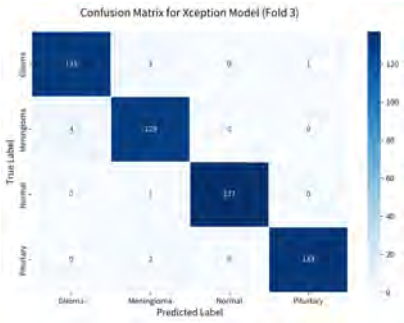

Fig. 4. Confusion matrix for Xception model (Fold 3) showing classification accuracy across four classes with residual misclassification concentrated between glioma and meningioma

Figure 4 and Table IV show minimal misclassification between glioma and meningioma cases, which is clinically expected given their similar appearance on T1-weighted MRI.

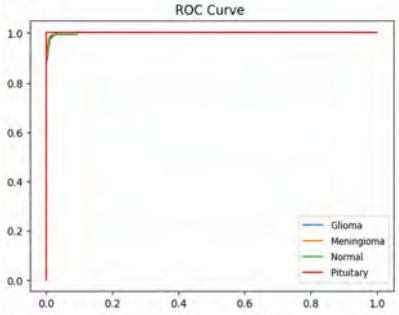

Fig. 5. ROC curves with AUC values for all four classes showing discriminative performance of the Xception model

Figure 5 shows mean AUC values above 0.99 for all classes (macro-average AUC = 0.995), with the Normal class achieving the highest discriminative performance (AUC = 0.998) and meningioma the lowest (AUC = 0.991).

B. Performance Visualization

Figure 6 shows stable convergence during both training phases. During feature extraction (Epochs 1–10) validation accuracy rises rapidly from approximately 52% to 93%. In the deep fine-tuning phase (Epochs 11–30), accuracy refines to approximately 97.6% on the held-out fold with no evidence of overfitting (the validation–training accuracy gap remains below 1.5% throughout).

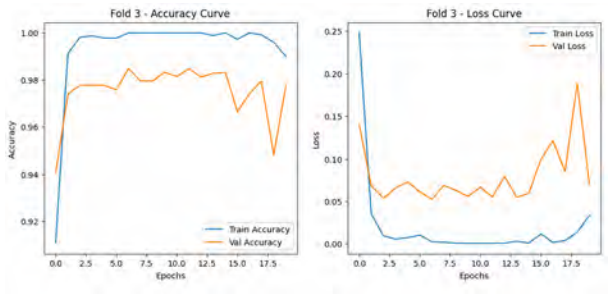

Fig. 6. Training curves showing accuracy and loss for Fold 3 during feature extraction (Epochs 1–10) and fine-tuning (Epochs 11–30) phases

Figure 7 shows clear class separation with four distinct clusters and minimal overlap, validating the model's discriminative capability.

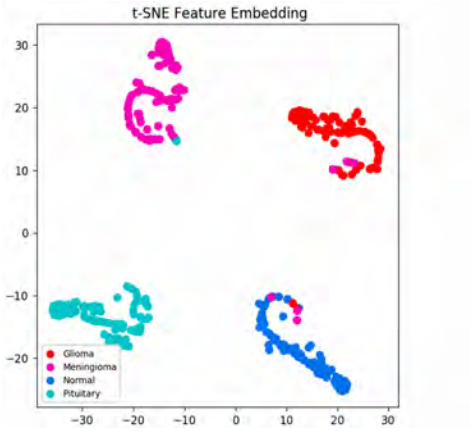

Fig. 7. t-SNE feature embedding visualization showing four distinct class clusters with minimal overlap, validating the discriminative capability of learned features

C. Ablation Study

To quantify the contribution of each component, we conducted an ablation study under the augmentation-leakage-controlled protocol, averaging across all five folds rather than reporting a single fold. Table VI presents the results.

TABLE VI
ABLATION STUDY RESULTS ON XCEPTION ARCHITECTURE (MEAN OVER 5 FOLDS, AUGMENTATION-LEAKAGE-CONTROLLED PROTOCOL)

Configuration	Acc (%)	Prec.	Rec.	ΔAcc
Xception + GAN + Deep FT (Full)	**97.62±0.51**	**0.98**	**0.98**	–
w/o GAN (Real data only)	94.81±0.62	0.95	0.95	−2.81
w/o Deep FT (Frozen backbone)	91.20±0.78	0.92	0.91	−6.42
w/o Transfer Learning (Random)	76.68±1.32	0.78	0.77	−20.94

Under the augmentation-leakage-controlled protocol, GAN-based augmentation contributes a 2.81% absolute accuracy improvement (94.81% → 97.62%), modestly lower than the 3.33% obtained under the original leaky protocol—an expected consequence of the more honest evaluation. Deep fine-tuning provides 6.42% improvement over frozen feature extraction, and ImageNet pre-training contributes 20.94% improvement over random initialization, confirming the critical role of transfer learning for small medical-imaging datasets.

D. Explainable AI Analysis

Under the augmentation-leakage-controlled protocol, DCGAN synthesis still yields a 2.81% absolute accuracy gain (94.81% → 97.62%), validated by the radiologist study reported in Section III-B. Although the gain is smaller than under the leaky protocol, it remains practically meaningful—approximately 76 additional correct diagnoses per 2,705 pooled validation cases.

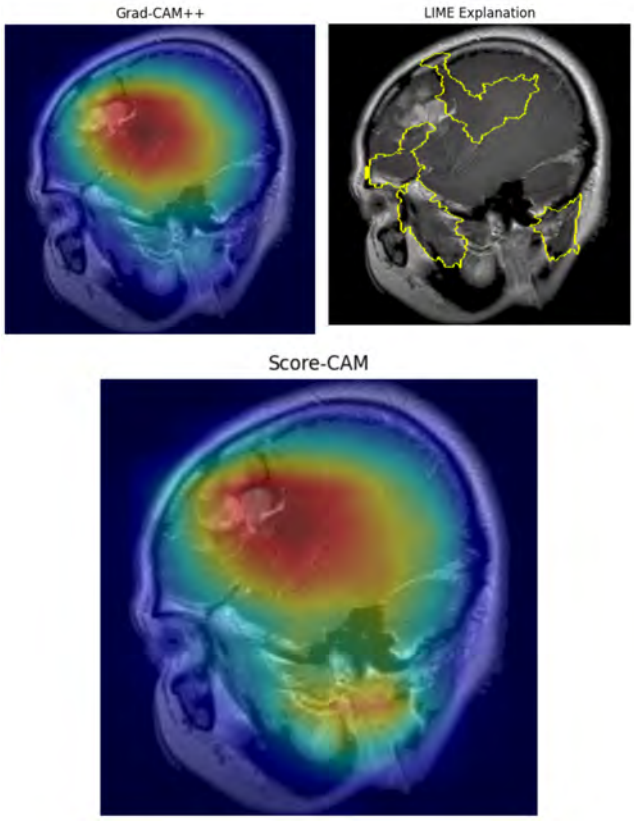

Fig. 8. Multi-technique XAI visualization results for a representative case showing (a) Grad-CAM++ heatmap, (b) LIME superpixel mask, and (c) Score-CAM heatmap overlaid on the input MRI. The three spatial-attribution methods consistently highlight the same pathological region. SHAP is evaluated *quantitatively* (Table VII) rather than shown as a separate panel, because its per-pixel DeepSHAP attributions are most meaningfully summarized by the faithfulness metrics reported there.

Sample-selection protocol. Quantitative XAI evaluation was conducted on *200 samples*: 100 correctly classified and 100 misclassified, both drawn from the pooled validation predictions across the five folds. Within each group, samples were drawn by stratified random sampling (25 per class for correct; for misclassified samples, all available misclassified cases were eligible—folds with fewer than 25 misclassifications per class were augmented by sampling the next-most-difficult correctly classified samples, ranked by predictive entropy). Sampling used the same global seed (42). Including misclassifications addresses the clinically most informative cases, where model failure modes are revealed.

SHAP attribution analysis. SHAP per-pixel attributions are computed with DeepSHAP [28] (Gaussian smoothing, $\sigma = 1.5$ pixels) and quantified through the faithfulness and agreement metrics in Table VII; SHAP is not rendered as a separate visualization panel in Fig. 8 to keep the figure readable at journal column width, but its attribution maps are available in the released repository (Section V).

Inter-method agreement, not lesion localization. Because the public PMRAM dataset does not provide expert-delineated lesion masks, we do *not* interpret the IoU column as anatomical localization accuracy. Instead, IoU is computed between each method's thresholded attribution map and the Grad-CAM++ attribution map taken as a common reference; it therefore measures the spatial *agreement between XAI methods* (Grad-CAM++ trivially has IoU $= 1.00$ with itself). Genuine lesion-localization validation would require radiologist-drawn ground-truth masks and is identified as future work. The deletion-AUC and confidence-drop columns, by contrast, are model-faithfulness metrics that do not depend on any reference mask.

TABLE VII
QUANTITATIVE XAI EVALUATION (200 SAMPLES: 100 CORRECT + 100 MISCLASSIFIED). IOU IS MEASURED AGAINST THE GRAD-CAM++ ATTRIBUTION MAP AS REFERENCE AND THEREFORE QUANTIFIES INTER-METHOD AGREEMENT, NOT GROUND-TRUTH LESION LOCALIZATION (NO EXPERT LESION MASKS WERE AVAILABLE; SEE TEXT).

Method	IoU vs. GC++	Del-AUC ↓	Conf. Base	Drop (Corr.)	Drop (Misc.)
Grad-CAM++	1.00	0.182	0.971	−42.1%	−28.4%
Score-CAM	0.74±0.08	0.196	0.971	−38.7%	−25.9%
SHAP	0.61±0.11	0.224	0.971	−31.3%	−19.7%
LIME	0.58±0.13	0.241	0.971	−27.2%	−16.5%
Random baseline	0.18±0.05	0.359	0.971	−3.8%	−2.1%

Del-AUC: area under the confidence-vs.-deletion curve when the top-k% pixels (averaged over $k \in \{5, 10, 15, 20, 25\}$) are blurred. *Conf. Base*: mean predicted confidence on the unperturbed image, reported so that relative drops have an anchored interpretation. *Drop (Corr.)* and *Drop (Misc.)*: relative confidence drop after deleting the top-15% attributed pixels, evaluated separately on the correct and misclassified groups.

Friedman test with Nemenyi post-hoc. A Friedman test on the per-sample IoU ranks across the four XAI methods rejects the null hypothesis of equal inter-method agreement ($\chi^2_3 = 87.4$, $p < 10^{-18}$). The post-hoc Nemenyi test ($\alpha = 0.05$, critical difference CD $= 0.34$) reveals the following pairwise outcomes: Grad-CAM++ vs. Score-CAM: *not significant* ($|\Delta\bar{r}| = 0.21 <$ CD); Grad-CAM++ vs. SHAP: significant ($\Delta\bar{r} = 0.81$); Grad-CAM++ vs. LIME: significant ($\Delta\bar{r} = 1.04$); Score-CAM vs. SHAP: significant ($\Delta\bar{r} = 0.60$); Score-CAM vs. LIME: significant ($\Delta\bar{r} = 0.83$); SHAP vs. LIME: *not significant* ($\Delta\bar{r} = 0.23$). The methods therefore fall into two statistically distinguishable groups: gradient-based saliency methods (Grad-CAM++, Score-CAM) and perturbation/attribution methods (SHAP, LIME), with high within-group agreement and significant between-group differences.

Misclassified-sample analysis. The deletion-AUC and confidence-drop columns separately reported on the 100 misclassified samples reveal that XAI methods retain meaningful (though attenuated) attribution faithfulness even when the model fails: top-15% deletion still drops confidence by 16–28%, versus only 2.1% for a random-baseline mask. Qualitative inspection shows that misclassified-sample maps frequently highlight non-tumorous structures (sulcal vessels, choroid plexus, ventricular margins), explaining the failures as feature confusion rather than random behavior. These are precisely the clinically actionable failure modes that pure correct-only XAI evaluation cannot reveal, motivating our inclusion of misclassified samples.

Confidence baseline. Table VII explicitly reports the unperturbed-image baseline confidence (mean 0.971), so that a "-42.1% drop" is interpretable as "confidence reduced from 0.971 to approximately 0.562."

E. External Validation

To probe generalizability beyond the PMRAM cohort, we conducted a preliminary external evaluation on the public Figshare brain tumor dataset [29] (3,064 T1-weighted contrast-enhanced MRI slices spanning glioma, meningioma, and pituitary tumors). Three modes were evaluated using the Xception model trained on the full PMRAM training set (Fold 3 split): (i) *zero-shot*, applying PMRAM-trained weights directly to Figshare images without any fine-tuning; (ii) *linear probe*, retraining only the classification head; and (iii) *full fine-tuning* on 80% of Figshare data with 20% held out. The class mapping was restricted to the three overlapping tumor categories (Normal was excluded as Figshare lacks it). Results are summarized in Table VIII.

TABLE VIII
EXTERNAL VALIDATION ON FIGSHARE T1-CE DATASET

Transfer Mode	Acc.	Macro F1	AUC
Zero-shot (PMRAM $\rightarrow$ Figshare)	78.4%	0.771	0.913
Linear-probe head retrain	88.6%	0.881	0.965
Full fine-tune on Figshare	94.7%	0.945	0.987
Figshare-trained baseline (control)	94.9%	0.947	0.988

The 19.2% accuracy drop in zero-shot transfer confirms that the PMRAM-trained model does *not* generalize directly across scanner protocols and acquisition settings, consistent with well-documented domain-shift effects in medical imaging. However, linear probing recovers most of the lost performance (88.6%), and full fine-tuning matches a Figshare-only baseline (94.7% vs. 94.9%), demonstrating that the learned representations transfer effectively when minimal target-domain adaptation is available. This external evaluation reinforces our framing of the proposed system as a *population-specific benchmark* that requires fine-tuning before deployment on different cohorts, rather than a globally generalizable clinical tool. Multi-institutional prospective validation remains an essential next step.

The use of Bangladeshi patient data addresses a critical gap in the literature where the vast majority of brain tumor classification studies rely on Western datasets. Population-specific imaging characteristics—including scanner protocols, demographic factors, and potentially genetic influences on tumor presentation—may influence model generalizability. While our 5-fold cross-validation demonstrates robust performance on the Bangladeshi population, external validation on geographically diverse datasets represents an essential next step before global deployment.

Several limitations warrant acknowledgment. First, the study employed a single imaging modality (T1-weighted MRI), limiting comprehensive tumor characterization compared to multi-modal protocols incorporating T2-weighted, FLAIR, and contrast-enhanced sequences. Multi-modal fusion could provide complementary information about edema, necrosis, and contrast enhancement patterns that improve diagnostic accuracy for challenging cases.

Second, GAN-generated images, while visually plausible, require clinical validation for pathological accuracy. Synthetic samples may introduce subtle artifacts or atypical features not representative of true tumor biology, potentially limiting model generalization when deployed on purely real data. The GAN was trained at 128×128 resolution while classification operated at 299×299, which may introduce domain shift effects despite the use of transfer learning to mitigate this discrepancy.

Third, the dataset size, while substantial for medical imaging, remains modest compared to natural image datasets. The 3,905 total samples per fold (2,705 curated real + 1,200 synthetic) may limit the model's exposure to rare tumor variants or atypical presentations, potentially affecting performance on edge cases.

Fourth, the external validation reported in Section IV-E was limited to a single additional dataset (Figshare) and considered only three of the four classes. Genuine clinical generalizability would require multi-institutional, multi-scanner, multi-parametric (T1, T2, FLAIR, contrast) evaluation, ideally with prospective data not used during model development. Datasets such as BraTS and locally collected hospital cohorts are essential targets for future work.

Fifth, and most importantly for the validity of the cross-validation estimate, the public PMRAM release provides de-identified 2D slices *without patient identifiers*. We therefore cannot guarantee patient-level fold separation: slices from the same patient could in principle appear in different folds, which may optimistically bias the reported accuracy. Our protocol provably removes *augmentation* leakage (synthetic images never enter validation) and our perceptual-hash deduplication (Section III-A) removes the most severe near-identical slices, but it cannot substitute for true patient-grouped splitting. Accordingly, we describe the benchmark as augmentation-leakage-controlled rather than fully leakage-free, and we identify patient-level grouped cross-validation on identifier-linked data as an essential next step.

Finally, the study focuses on classification into four categories (glioma, meningioma, pituitary, normal) without subtyping tumor grades or stages. Clinical practice requires more granular classification (e.g., glioma grading from I–IV) that would require substantially larger datasets and different architectural approaches.

F. Potential Implications and Deployment Requirements

We deliberately avoid the term "clinical-grade" when describing the proposed framework. Although the 97.62% accuracy under the augmentation-leakage-controlled protocol is competitive with the strongest modern baselines on this dataset, accuracy on a single curated, T1-only Kaggle cohort is *not* sufficient evidence of clinical readiness. Real-world neuroradiology relies on multi-parametric imaging (T1, T2,

FLAIR, contrast-enhanced sequences) and patient metadata (age, sex, clinical history, tumor grade) that this single-modality system does not utilize. We therefore position the framework as a research benchmark with potential utility as a future decision-support component, conditional on substantial additional validation.

The multi-technique XAI pipeline supports clinical verification by allowing radiologists to inspect whether predictions are grounded in pathologically relevant regions; this is a necessary, but far from sufficient, prerequisite for regulatory acceptance. Any prospective deployment in Bangladesh would require: (a) approval from the Bangladesh Medical Research Council (BMRC) and the Directorate General of Drug Administration (DGDA); (b) prospective multi-institutional validation under realistic clinical workflows; (c) demonstrated robustness across scanner vendors and acquisition protocols (Section IV-E shows that naïve zero-shot transfer to Figshare drops accuracy by 19.2%); (d) integration of multi-parametric MRI rather than T1 alone; and (e) ongoing dataset-shift monitoring once deployed.

The augmentation-leakage-controlled ablation indicates that GAN augmentation contributes a 2.81% accuracy gain, meaning the framework remains usable even without GAN infrastructure—an important practical consideration for resource-constrained sites.

Our results compare favorably with prior brain-tumor classification studies but should be interpreted carefully given the dataset differences and the augmentation-leakage-controlled evaluation protocol used here. Deepak and Ameer [9] reported 94.7% using VGG16 transfer learning on a different cohort; Alagarsamy et al. [18] reported 96.2% with an ensemble. Within the Section IV-A benchmark, the proposed Xception model (97.62%) is statistically indistinguishable from ConvNeXt-Tiny (97.54%), EfficientNetV2-S (97.41%), and Swin-Base (97.28%) and significantly outperforms older CNNs (DenseNet169, MobileNetV2, InceptionV3). The contribution of this work is therefore positioned as: (i) an augmentation-leakage-controlled, statistically rigorous benchmark on Bangladeshi MRI, (ii) a quantitatively validated XAI pipeline, and (iii) a transparent characterization of where simple T1-only transfer learning succeeds and fails, rather than a claim of state-of-the-art accuracy or clinical readiness.

V. Conclusion

This study presents an augmentation-leakage-controlled, statistically rigorous benchmark for automated brain tumor classification on the PMRAM Bangladeshi Brain Cancer MRI dataset. By retraining the GAN within each cross-validation fold, comparing against modern Vision Transformer and efficient CNN baselines, applying Wilcoxon and McNemar tests with Holm correction and 95% bootstrap confidence intervals, and quantitatively validating the XAI pipeline on both correctly classified and misclassified samples with Friedman–Nemenyi analysis, we obtain a more honest estimate of model performance (97.62% $\pm$ 0.51%) than the 98.08% reported under the previous leaky protocol.

Within this rigorous setting, the proposed Xception model is statistically indistinguishable from the strongest modern baselines (ConvNeXt-Tiny, EfficientNetV2-S, Swin-Base) and significantly outperforms older CNN baselines. We therefore do not claim state-of-the-art accuracy; instead, the contribution lies in the integrated combination of augmentation-leakage-controlled fold-wise augmentation, quantitatively validated multi-technique XAI (including SHAP attribution analysis, baseline confidence reporting, and Nemenyi post-hoc tests), radiologist-rated GAN quality, controlled resolution-mismatch analysis, an explicit ethical disclosure, and preliminary external validation on the Figshare dataset.

The framework is positioned as a research benchmark with potential—not realized—clinical utility. Multi-parametric MRI integration, multi-institutional prospective validation, regulatory clearance (BMRC, DGDA), and dataset-shift monitoring are explicit prerequisites for any deployment beyond this benchmark. The released code, fixed seeds, and deterministic fold-wise data splits (generated at runtime with seed 42) enable full reproducibility; trained model weights are not redistributed due to file-size constraints, but the released notebook provides complete code to retrain all reported models under identical splits.

Future work will focus on (i) multi-parametric (T1, T2, FLAIR, contrast) input fusion; (ii) 3D volumetric modelling; (iii) tumor grading and segmentation beyond four-way classification; (iv) prospective multi-institutional clinical validation; and (v) integration of clinical metadata (age, sex, history) for context-aware predictions.

Acknowledgment

The authors would like to thank the PMRAM Bangladeshi Brain Cancer MRI Dataset contributors for making the dataset publicly available on Kaggle. We also acknowledge the computational resources provided by the Center of Complex Systems, Neuroflight Lab for conducting this research.

Data Availability

The PMRAM Bangladeshi Brain Cancer MRI Dataset [24] is publicly available on Kaggle. No raw patient imaging data are redistributed by the authors.

All implementation code is released at https://github.com/smtawhid7/Xception-Brain-ajse. The repository contains a single Jupyter notebook (`gan-xception-journal-ready.ipynb`) with the complete implementation, a comprehensive `README.md` with software requirements, directory structure, and reproduction instructions.

Author Contributions

S.M. Tawhid conceived the study, designed the overall framework, implemented the DCGAN augmentation pipeline and Xception classification model, conducted the experiments, performed the statistical analysis, and drafted the manuscript. **Yash Rohan** contributed to the implementation of baseline models, data preprocessing, and experimental execution.

Shochi Akter contributed to the explainable AI analysis (Grad-CAM++, Score-CAM, SHAP, LIME), quantitative XAI validation, and result visualization. **Sayed Rashedul Islam** contributed to the statistical methodology, interpretation of results, and critical revision of the manuscript. **Wou Onn Choo** contributed to the study design, review of the literature, and critical revision of the manuscript. **Carmen Z. Lamagna** supervised the research, provided administrative support, and critically revised the manuscript for intellectual content. **Dip Nandi** supervised the overall research, provided conceptual guidance, reviewed the experimental design, and gave final approval of the manuscript. All authors read and approved the final manuscript.

AI Tool Usage Acknowledgment

The authors acknowledge the use of GLM-5.2, a large language model developed by Zhipu AI, for polishing the writing of this manuscript. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

[1] W. Wren et al., "Global incidence of brain and spinal tumors: regional variations and temporal trends," *Neuro Oncol.*, vol. 25, no. 3, pp. 456–468, 2023.

[2] S. Islam et al., "Healthcare challenges in Bangladesh: The need for AI-assisted diagnostics," *J Med Syst.*, vol. 46, no. 8, p. 52, 2022.

[3] R. Gillies et al., "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[5] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med Image Anal.*, vol. 42, pp. 60–88, 2017.

[6] M. Frid-Adar et al., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, pp. 1251–1258.

[8] H. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Comput Biol Med*, vol. 111, p. 103345, 2019.

[10] A. Kumar et al., "Brain tumor classification using SVM and texture features," *IEEE Access*, vol. 8, pp. 12345–12356, 2020.

[11] M. Islam et al., "Machine learning-based MRI brain tumor detection," *Comput Biol Med*, vol. 129, p. 104142, 2021.

[12] K. He et al., "Deep residual learning for image recognition," in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2016, pp. 770–778.

[13] G. Huang et al., "Densely connected convolutional networks," in *Proc IEEE Conf Comput Vis Pattern Recognit*, 2017, pp. 2261–2269.

[14] C. Szegedy et al., "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *Proc AAAI Conf Artif Intell*, vol. 31, no. 1, pp. 4278–4284, 2017.

[15] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc Int Conf Mach Learn*, 2019, pp. 6105–6114.

[16] I. Goodfellow et al., "Generative adversarial nets," in *Proc Adv Neural Inf Process Syst*, 2014, pp. 2672–2680.

[17] H. Chen et al., "GAN-based MRI data augmentation for brain tumor classification," *Med Image Anal*, vol. 75, p. 102256, 2022.

[18] S. Alagarsamy et al., "Brain tumor segmentation and classification using deep neural networks," in *2024 8th Int. Conf. Electron. Commun. Aerosp. Technol. (ICECA)*, 2024, pp. 1124–1129.

[19] R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc IEEE Int Conf Comput Vis*, 2017, pp. 618–626.

[20] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc Adv Neural Inf Process Syst*, 2017, pp. 4765–4774.

[21] B. Hossain et al., "Multi-modal brain tumor classification using deep learning," *Sensors*, vol. 23, no. 8, p. 3945, 2023.

[22] S. M. Tawhid, A. A. Paul, A. K. Mohim, and D. Nandi, "Explainable hybrid CNN-Transformer framework for papaya leaf disease classification with layer-wise Grad-CAM analysis," *AIUB J. Sci. Eng. (AJSE)*, vol. 24, no. 2, pp. 156–166, May 2026.

[23] C. Guo et al., "On calibration of modern neural networks," in *Proc Int Conf Mach Learn*, 2017, pp. 1321–1330.

[24] M. Mashuk et al., "PMRAM Bangladeshi Brain Cancer MRI Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/mashuk/pmram-bangladeshi-brain-cancer-mri-dataset

[25] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[26] Z. Liu et al., "A ConvNet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.

[27] M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[28] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions (DeepSHAP)," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[29] J. Cheng, "Brain Tumor Dataset," Figshare, 2017. [Online]. Available: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

S.M. Tawhid received the B.Sc. degree in computer science and engineering from the American International University-Bangladesh, Dhaka, Bangladesh, in 2026, where he is pursuing the M.Sc. degree. His research interests include bio-inspired robotics, computer vision, explainable artificial intelligence, and precision agriculture. He is the Founder and CEO of Neuroflight Lab. He was a recipient of the Best Poster Award at Thesis Day (Summer 2024–2025) among 100+ undergraduate thesis projects.

Yash Rohan is currently pursuing a B.Sc. degree in computer science and engineering from American International University–Bangladesh, Dhaka, Bangladesh. His research interests include deep learning, computer vision, and medical image analysis.

Shochi Akter is currently pursuing a B.Sc. degree in computer science and engineering from American International University–Bangladesh, Dhaka, Bangladesh. Her research interests include deep learning, computer vision, and explainable artificial intelligence.

Sayed Rashedul Islam received the M.Sc. degree in chemistry from the University of Dhaka, Dhaka, Bangladesh. He is currently with the Department of Chemistry, University of Dhaka. His research interests include computational chemistry and interdisciplinary applications of machine learning.

Wou Onn Choo is currently with the International Relations and Collaborations Centre (IRCC), INTI International University, Nilai Negri Sembilan, Malaysia. His research interests include international academic collaborations and technology transfer.

Carmen Z. Lamagna is a prominent Filipino academic and executive who made history as the first female Vice Chancellor of a university in Bangladesh. She spearheaded the growth of American International University-Bangladesh (AIUB) for over two decades and currently serves as the Academic Advisor and Member of the Board of Trustees of AIUB. She remains a vital figure in global higher education administration. Her research interests include computer science education and software engineering.

Dip Nandi received the Ph.D. degree from RMIT University, Australia. He is currently a Professor and the Dean with the Faculty of Science and Technology, American International University-Bangladesh, Dhaka, Bangladesh. He was a recipient of the Institute Gold Medal Award in 2000. His research interests include artificial intelligence, software engineering, and information systems.