CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Journal Paper

Efficient and Explainable MRI Brain Tumor Classification via Adaptive GAN-Based Data Augmentation

Anbazhagan Mahadevan
Radha Reddy*

*CISTER Research Centre
CISTER-TR-251103

2025/11

Efficient and Explainable MRI Brain Tumor Classification via Adaptive GAN-Based Data Augmentation

Anbazhagan Mahadevan, Radha Reddy*

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: mailanbazhagan@gmail.com, reddy@isep.ipp.pt

https://www.cister-labs.pt

Abstract

Accurate brain tumor diagnosis from MRI is critical, but state-of-the-art deep learning (DL) models are computationally expensive and require large datasets, limiting clinical deployment. Classical machine learning (ML) is more efficient but struggles with imbalanced data and a persistent performance gap. This study proposes an XAI-driven Adaptive GAN (AGAN) framework to bridge this gap. The AGAN integrates interpretability scores (e.g., SHAP, Grad-CAM) into its training objective to iteratively refine synthetic sample generation, focusing on improving minority tumor class representation. Evaluated on four diverse public MRI datasets, the AGAN framework demonstrated a 19.7% improvement in minority-class F1-scoreover the baseline. AGAN-augmented classical ML classifiers, particularly Support Vector Machine (SVM), achieved accuracies up to 0.902, narrowing the performance gap to state-of-the-art DL models (e.g., ResNet50) to within 2.8 135.7%. Critically, the ML pipeline delivered up to 7000× faster training and required up to 240× less memory. These findings validate the XAI-guided, AGAN-augmented ML pipeline as a scalable, resource-efficient, and interpretable alternative for brain tumor diagnosis, highly suitable fordeployment in resource-constrained clinical environments.

Array 28 (2025) 100609

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Efficient and explainable MRI brain tumor classification via Adaptive GAN-based data augmentation

Anbazhagan Mahadevan [a], Radha Reddy [b,c,*], Ujwal Nadimpalli [a], Mahammad Khaji [a], Bhavana Nannuru [a], Aateesh Akkina [a]

[a] Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, Amrita Vishwa Vidyapeetham, India
[b] Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India
[c] CISTER - Research Centre in Real-Time and Embedded Computing Systems, Porto, 4200-135, Portugal

ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri, https://figshare.com/articles/dataset/brain_tumor_dataset/1512427, https://www.kaggle.com/datasets/mohammadhossein77/brain-tumors-dataset, https://www.kaggle.com/datasets/rishiksaisanthosh/brain-tumour-classification

Keywords:
Brain tumor classification
Dimensionality reduction
Explainable AI
Generative Adversarial Networks
Machine learning

ABSTRACT

Accurate brain tumor diagnosis from MRI is critical, but state-of-the-art deep learning (DL) models are computationally expensive and require large datasets, limiting clinical deployment. Classical machine learning (ML) is more efficient but struggles with imbalanced data and a persistent performance gap. This study proposes an XAI-driven Adaptive GAN (AGAN) framework to bridge this gap. The AGAN integrates interpretability scores (e.g., SHAP, Grad-CAM) into its training objective to iteratively refine synthetic sample generation, focusing on improving minority tumor class representation. Evaluated on four diverse public MRI datasets, the AGAN framework demonstrated a 19.7% improvement in minority-class F1-score over the baseline. AGAN-augmented classical ML classifiers, particularly Support Vector Machine (SVM), achieved accuracies up to 0.902, narrowing the performance gap to state-of-the-art DL models (e.g., ResNet50) to within 2.8–5.7%. Critically, the ML pipeline delivered up to 7000× faster training and required up to 240× less memory. These findings validate the XAI-guided, AGAN-augmented ML pipeline as a scalable, resource-efficient, and interpretable alternative for brain tumor diagnosis, highly suitable for deployment in resource-constrained clinical environments.

1. Introduction

Tumors and abnormal growths within or around the brain substantially threaten human health, frequently disrupting neurological functions such as memory, cognition, speech, and motor skills [1]. Early and accurate diagnosis is paramount. Misclassifications or diagnostic delays can negatively affect treatment outcomes, patient survival, and quality of life. The MRI scans remain the gold standard for non-invasive tumor assessment, providing detailed, high-resolution anatomical views across multiple planes and tumor types.

Despite the advantages of imaging techniques, the reliable classification of brain tumors from MRI remains a significant clinical challenge. Tumors exhibit significant heterogeneity in shape, size, anatomical location, and MRI appearance, complicating their prompt and accurate diagnosis. Fig. 1, adapted from [2], visually underscores this diversity by presenting sample MRI scans of both normal brains and multiple tumor classes commonly encountered in practice. Such variability underscores the need for automated, robust classification methods to support consistent, accurate clinical decision-making.

ML and DL have recently transformed medical image analysis by enabling the automated detection of subtle imaging patterns, often imperceptible to human radiologists [3–5]. DL models, particularly CNNs, have successfully been used for brain tumor identification, treatment planning, and outcome prediction. However, these approaches typically rely on complex model architectures, large volumes of expertly annotated data, and considerable computational resources [6]. These requirements limit their deployment in resource-constrained scenarios, such as rural clinics or embedded healthcare devices.

Conversely, classical ML models, such as Decision Trees (DT), K-Nearest Neighbors (KNN), and SVM, possess significant, but underexplored, potential areas [7]. With optimized preprocessing and dimensionality reduction, these models can offer competitive predictive accuracy while requiring considerably less computational overhead [8].

Given their high spatial resolution, MRI scans produce large pixel-intensity matrices that encode rich structural and pathophysiological information [9,10]. However, this high dimensionality introduces analytical challenges, most notably the "curse of dimensionality", which increases the risk of overfitting and compromises model generalization, especially on small or imbalanced datasets [3]. Another persistent

* Corresponding author at: Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India.
E-mail address: radhakrishna.reddy@manipal.edu (R. Reddy).

https://doi.org/10.1016/j.array.2025.100609
Received 7 September 2025; Received in revised form 19 November 2025; Accepted 25 November 2025
Available online 29 November 2025

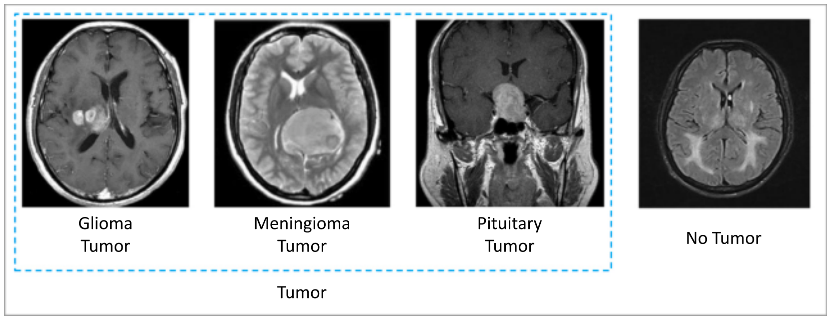


Fig. 1. Sample MRI scan images depicting normal brain and different types of brain tumors [2].

challenge is severe class imbalance among tumor types; models tend to overfit to the majority classes, reducing sensitivity to clinically crucial minority groups [11].

To address these gaps, this study constructs structured, numerically encoded representations of MRI scans by leveraging pixel-level RGB intensities, ensuring compatibility with both ML and DL algorithms. Principal component analysis (PCA) is employed to retain critical information while minimizing computational load and enhancing interpretability [12].

An adaptive Generative Adversarial Network (GAN) framework is a key methodological innovation that produces synthetic samples specifically for underrepresented tumor classes. Traditional GAN-based data augmentation typically generates samples uniformly across all classes. In contrast, our adaptive strategy prioritizes classes with limited data availability, guided by both class distribution statistics and interpretability feedback from the model. This tailored augmentation improves model balancing and discriminative power and supports efficient deployment in constrained clinical environments.

Interpretability remains essential for clinical applications, where trust and transparency are prerequisites for adoption alongside accuracy. To this end, we embed state-of-the-art explainable AI (XAI) techniques, including SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM) [13]. These methods generate local and global explanations, illuminating which image features or regions most strongly influence model predictions and bridging the critical gap between black-box algorithms and clinical reasoning.

In summary, this work directly responds to persistent challenges in brain tumor classification by:

- A thorough comparative evaluation across classical ML and DL models on diverse MRI datasets highlights the strengths and limitations of each approach.
- The lightweight preprocessing pipeline, with structured feature encoding and PCA-based dimensionality reduction, ensures efficient model training without sacrificing critical diagnostic information.
- Developing an AGAN that selectively targets underrepresented tumor classes, improving class balance and boosting minority-class performance through an XAI-driven feedback loop.
- Embedding explainable AI methods (SHAP, LIME, Grad-CAM) provides transparent, clinically meaningful insights into model decisions, fostering trust and supporting adoption.
- Validating the framework's performance, efficiency, and interpretability on four public datasets, demonstrating real-world applicability and resource-efficient scalability.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature and defines the research gaps. Section 3 outlines the proposed methodology, and mathematical formulation and algorithmic framework for the proposed AGAN strategy. Section 4 describes data acquisition and preparation. Section 5 discusses XAI techniques for model interpretability. Section 6 presents experimental results and comparative performance analyses. Section 7 provides a focused interpretability assessment. Section 8 summarizes key findings and limitations. Finally, Section 9 concludes with implications and future research directions.

2. Related works

Brain tumor classification has attracted substantial research attention, with numerous studies leveraging both traditional ML and advanced DL methodologies for medical image analysis. Early efforts focused on classical ML models, such as SVMs and DTs, that require manual feature extraction and often yield reliable, interpretable results. The rise of DL, especially CNNs, enabled automated feature learning and notable accuracy improvements, prompting the frequent deployment of architectures such as VGG, ResNet, and hybrid ensembles in recent years.

2.1. Machine learning techniques

Early and ongoing ML approaches, including SVM, Naïve Bayes (NB), DTs, and various ensemble classifiers, have successfully classified brain tumors from MRI and CT images using extracted features. For example, Solanki et al. [14] reported SVM as the top classical classifier, though deep models like VGG16 outperformed it, achieving 97.86% accuracy. The value of dimensionality reduction (e.g., PCA) and powerful ensemble methods such as XGBoost (XGB) is highlighted by Kuppusamy et al. [15], albeit with increased computational cost. Hybrid and ensemble techniques further improved reliability, as seen in [16,17], where combining SVM with NB or fusing multiple pretrained models achieved up to 99% accuracy. Comparative studies, such as [18], demonstrate that hybrid models can even match or surpass deep models, underlining the ongoing relevance of optimized ML techniques, particularly for their interpretability and lower resource requirements.

2.2. Deep learning advancements

Developing sophisticated DL models, notably EfficientNet, Inception-ResNetV2, and transfer learning architectures, has further boosted performance in brain tumor classification. Models like EfficientNet achieved up to 99.82% training and 98.78% testing accuracy [19], while Inception-ResNetV2 in [20] reached 99.98%. Hybrid approaches are increasingly common: feature extraction by ResNet-18 and GoogLeNet paired with SVM reached 98% accuracy [21], and CNN-KNN hybrids, as in [22], attained 99.45%. Adding texture-based features (GLCM) and transfer learning methods has enhanced robustness across datasets [23], with the integration of ML and DL often outperforming standalone strategies.

2.3. GANs in medical imaging

GANs are now central to tackling class imbalance and dataset limitations in brain tumor imaging. Advanced GANs, such as Uncertainty-Aware GANs and class-focused techniques, improve both MRI enhancement and sample diversity [24]. Hybrid noise-to-image GANs have increased structural realism and boosted detection rates [25]. Segmentation networks like DeepMRSeg, augmented with GAN-generated images, achieved 91.7% accuracy [26], while TumorGANet [27] and DIR-GAN [28] achieved near-perfect results by leveraging augmentation and explainability. These studies demonstrate GANs' vital role in addressing class imbalance and improving the validity of AI-driven imaging.

2.4. Explainable AI

Integrating XAI has become essential for clinical adoption and trust, particularly as DL models are often viewed as "black boxes". Recent research has developed CNNs enhanced with SHAP, LIME, and Grad-CAM, improving transparency and interpretability. For instance, Nazir et al. [29] combined a custom CNN with these XAI tools, achieving 98.67% accuracy and detailed visual explanations. Gaur et al. [30] extended this using SHAP and LIME with a dual-input CNN robust to noise. LIME and Grad-CAM have further empowered VGG16, ResNet50, and DenseNet models to achieve high accuracy and simultaneously provide actionable visual attribution, as demonstrated in [31,32]. Together, these advances highlight the clinical demands for reliability, explanation, and human validation in AI-based neuroimaging.

2.5. Summary

While DL and advanced ensemble methods have led to recent progress, much of the field still focuses on binary classification and complex architectures, often lacking interpretability or direct applicability to diverse clinical cases. Few studies take on rigorous multi-class classification with explicit class-aware augmentation and integrated explainability. As such, the gaps addressed by this study include: the underutilization of classical ML for multi-class tasks; insufficient consideration of class balance through adaptive augmentation; and limited adoption of transparent, interpretable frameworks to support clinical decision-making. Our work responds to these gaps by combining robust classical ML models, targeted GAN-based data synthesis, and embedded XAI, providing a unified, clinically practical solution for brain tumor classification.

3. Proposed methodology

This section presents a proposed, structured, and generalizable image classification pipeline that combines classical ML models with AGAN-based data augmentation and post-hoc XAI techniques. This framework addresses class imbalance and reduces feature dimensionality through iterative PCA. It integrates interpretability tools such as SHAP, LIME, and Grad-CAM. Unlike DL-centric workflows, this approach demonstrates that classical ML, coupled with principled data balancing and interpretability modules, can achieve competitive performance with improved computational efficiency and transparency. The proposed workflow, illustrated in Fig. 2, is modular and systematic, comprising seven key stages.

The workflow begins with **Step 1: Image Preprocessing**, where raw brain MRI scans are standardized to ensure input consistency. This involves resizing images to a uniform 224×224 dimension, converting grayscale images to RGB format where necessary, and applying pixel-wise normalization. These preprocessing steps are essential to eliminate noise and discrepancies introduced by varying imaging conditions and acquisition protocols across datasets.

Step 2: Imbalance Handling evaluates whether the dataset exhibits class imbalance, a common challenge in medical datasets where certain tumor types are significantly underrepresented. When an imbalance is detected, the system proceeds to AGAN Synthesis, which selectively trains class-specific GANs to generate high-quality synthetic samples for minority tumor categories. This adaptive approach integrates XAI feedback, allowing the GAN to iteratively refine its generation process based on interpretability analyses of previously generated samples (only synthetic samples were considered to prevent test data leakage). The process continues until adequate representation is achieved for all minority classes.

Step 3: Data Preparation combines the original dataset with newly generated synthetic samples, ensuring a balanced representation across all tumor classes. Additional traditional augmentation techniques, including rotation, flipping, brightness variation, and noise injection, are applied to further enhance sample diversity. The augmented dataset is then split into training and validation sets using stratified sampling to maintain class proportions.

Step 4: Dimensionality Reduction addresses the computational challenges posed by high-dimensional MRI features. PCA identifies components that retain 95% of the original variance while significantly reducing feature dimensionality. The process includes systematic component analysis and feature optimization, with iterative evaluation to determine the optimal number of components that balance computational efficiency with discriminative power.

Step 5: Model Training and Selection encompasses comprehensive model development and evaluation. Multiple classical ML algorithms, including SVM, KNN, DTs, and ensemble methods (RF & XGB), undergo hyperparameter tuning using randomized search strategies. Cross-validation techniques ensure robust performance estimation, followed by systematic performance evaluation across multiple metrics. The best-performing model is selected based on accuracy, computational efficiency, and clinical applicability.

Step 6: XAI Interpretation enhances model transparency through multiple explainability techniques. SHAP analysis provides global feature attribution, revealing which pixel regions considerably influence predictions across different tumor classes. LIME generates local explanations for individual predictions, highlighting specific image regions that support or oppose classification decisions. For DL models, Grad-CAM produces attention maps that highlight class-discriminative regions in input images. Importantly, the insights gained from these XAI techniques provide feedback to refine the AGAN synthesis process, creating a closed-loop system that continuously improves synthetic sample quality.

Finally, **Step 7: Final Output** delivers comprehensive tumor classification results accompanied by clinical explanations. The system outputs the predicted tumor class and detailed visual and textual explanations derived from the XAI analysis. Combining accurate predictions with interpretable reasoning, this dual output facilitates clinical validation and supports informed decision-making in diagnostic workflows.

3.1. AGAN-based class imbalance handling

Class imbalance is a major challenge in medical image classification, especially when rare tumor types are underrepresented. Standard ML models often favor majority classes to maximize overall accuracy, leading to poor recall and precision for minority classes. This is critical in medical diagnostics, where false negatives, such as misclassifying a tumor as "no tumor", can be life-threatening.

To address this issue, we propose an AGAN framework that dynamically adjusts the number and diversity of synthetic samples for each minority class based on class distribution and model interpretability feedback. Unlike traditional GANs or oversampling methods such as Synthetic Minority Oversampling TEchnique (SMOTE), our AGAN uses class-specific generation tuning enhanced with XAI-guided quality

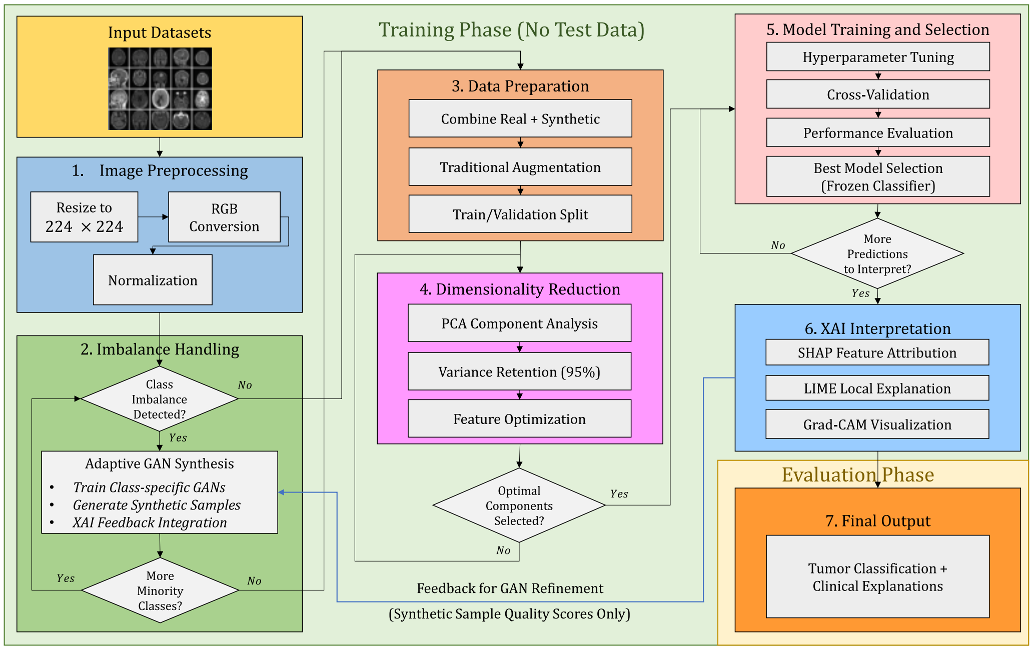


Fig. 2. Work flow diagram of the proposed approach.

control to ensure that each minority class receives targeted and relevant synthetic enhancement. The key innovation lies in the iterative feedback loop where XAI interpretability metrics guide the refinement of synthetic sample generation. All mathematical symbols and abbreviations employed in this work are defined in Table A.8.

(1) XAI-Guided Quality Assessment: For each synthetic sample x_i generated by the GAN, we compute an XAI quality score $Q_{\mathrm{XAI}}(x_i)$ that quantifies the interpretability and clinical relevance of the synthetic sample. This score can be formulated using different XAI metrics:

For Grad-CAM based assessment (Eq. (1)):

$$Q_{\mathrm{XAI}}(x_i) = \mathrm{IoU}(\mathrm{GradCAM}(x_i), \mathrm{TumorMask}(x_i)) \tag{1}$$

For SHAP-based feature attribution (Eq. (2)):

$$Q_{\mathrm{XAI}}(x_i) = \sum_{j \in \text{tumor-region}} |\phi_j(\mathrm{SHAP}(x_i))| \tag{2}$$

where ϕ_j represents the SHAP value for pixel/feature j in the tumor region.

(2) Adaptive Generator with Feedback Integration: The generator G_c for class c is enhanced to incorporate XAI feedback through a modified training objective (Eq. (3)):

$$G_c(z) = f(W_g^{(c)} z + b_g^{(c)}) \tag{3}$$

where $W_g^{(c)}$ and $b_g^{(c)}$ are class-specific learned parameters, and f is a non-linear activation function.

(3) XAI-Enhanced Discriminator: The discriminator D evaluates both the realism and interpretability of generated samples (Eq. (4)):

$$D(x) = \sigma(W_d x + b_d) \tag{4}$$

where σ is the sigmoid activation, and W_d, b_d are the discriminator parameters.

(4) Modified Loss Function with XAI Penalty: The training objective integrates XAI feedback through an interpretability penalty term (Eq. (5)):

$$\begin{aligned}\min_G \max_D V(D,G) = \sum_{c \in C} \lambda_c \Big[& \mathbb{E}_{x \sim p_{data}^c}[\log D(x)] \\ & + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G_c(z)))] \\ & + \lambda_{\mathrm{XAI}} \mathbb{E}_{x_i \sim G_c}[1 - Q_{\mathrm{XAI}}(x_i)] \Big]\end{aligned} \tag{5}$$

where λ_{XAI} controls the influence of interpretability feedback, and the imbalance penalty λ_c is (Eq. (6)):

$$\lambda_c = \frac{1}{n_c} \bigg/ \sum_{j=1}^{|C|} \frac{1}{n_j} \tag{6}$$

(5) Iterative Refinement Process: The AGAN operates through iterative feedback cycles. At each iteration t, synthetic samples are generated, evaluated using XAI metrics, and filtered based on a quality threshold τ_{XAI} (Eq. (7)):

$$X_{\text{filtered},c}^{(t)} = \{x_i \mid Q_{\mathrm{XAI}}(x_i) \geq \tau_{\mathrm{XAI}}\} \tag{7}$$

Only samples that meet the interpretability threshold are retained for subsequent GAN training iterations, ensuring a progressive improvement in synthetic sample quality. The selection of the XAI quality threshold, τ_{XAI}, follows a systematic and performance-driven strategy aimed at maximizing the downstream classifier's effectiveness without the need for external annotations. Specifically, we conducted a comprehensive grid search over threshold values $\tau \in [0.3, 0.9]$ with an increment of 0.05 (the representative values are shown in Table 1). For each candidate threshold: (i) Generate synthetic samples using AGAN, (ii) Filter samples: $\mathcal{X}_{\text{filtered}} = \{x_i \mid Q_{\mathrm{XAI}}(x_i) \geq \tau\}$, (iii) Train classifiers on augmented dataset: $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{original}} \cup \mathcal{X}_{\text{filtered}}$, and (iv) Evaluate using 5-fold cross-validation. The optimal threshold is maximized using the following expression (Eq. (8)):

$$\begin{aligned}\tau_{\mathrm{XAI}}^* = \arg\max_{\tau} [0.5 \cdot F1_{\text{minority}}(\tau) + 0.3 \cdot \mathrm{Accuracy}(\tau) \\ + 0.2 \cdot \mathrm{RetentionRate}(\tau)]\end{aligned} \tag{8}$$

This weighting scheme prioritizes performance improvement in minority classes, our primary objective, while preserving overall model performance and sample diversity.

The selected thresholds were required to satisfy the following criteria: (i) minority class F1-score improvement of at least 10% over the baseline, (ii) overall accuracy degradation not exceeding 2%, (iii) a retention rate of at least 25% to ensure adequate sample diversity, and (iv) consistent performance across all four datasets. Following this evaluation, $\tau_{\mathrm{XAI}} = 0.60$ was identified as the optimal threshold, resulting in a 15% improvement in minority class F1-score while maintaining 89% overall accuracy and a 33% retention rate across datasets.

Table 1
Threshold selection results using performance-driven optimization.

Threshold	Minority F1	Overall Acc	Retention rate	Composite score
0.45	0.72 ± 0.06	0.86 ± 0.04	45.2%	0.681
0.55	0.76 ± 0.05	0.88 ± 0.03	36.8%	0.724
0.60	**0.78 ± 0.04**	**0.89 ± 0.03**	**33.1%**	**0.741**
0.65	0.77 ± 0.05	0.89 ± 0.03	28.5%	0.715
0.75	0.74 ± 0.06	0.88 ± 0.04	18.9%	0.656

(6) Data Leakage Prevention: Our XAI-guided feedback mechanism is designed with multiple safeguards to prevent test data leakage, ensuring that no test set information influences synthetic sample generation:

Temporal Separation Protocol:

1. **Phase 1:** The baseline classifier is trained exclusively on the training and validation splits, which correspond to 80% and 20% of the original dataset, respectively.
2. **Phase 2:** XAI evaluation operates only on synthetic samples generated by the GAN, but not on actual test data.
3. **Phase 3:** Final model evaluation uses completely independent test sets (the remaining 20%) that are never accessed during GAN training.

Information Flow Control: The workflow ensures strict information isolation through a frozen baseline classifier architecture. XAI interpretability scores are computed exclusively on synthetic samples using this frozen classifier, with no gradients or parameter updates flowing back from XAI evaluation to the baseline model. Test data remains completely isolated until the final evaluation phase. During threshold optimization, cross-validation is performed solely on the training and validation splits, thereby maintaining complete separation from all test instances and preventing information leakage.

Algorithmic Safeguards: The framework implements multiple computational safeguards to prevent data leakage. The baseline classifier parameters are permanently frozen after initial training, ensuring that XAI feedback influences only the GAN generator parameters without affecting the evaluation classifier. Synthetic sample filtering is performed in a separate computational loop without access to the test data. The final augmented dataset is constructed before any test set evaluation, establishing clear temporal boundaries between the training and evaluation phases.

(7) Final Augmented Dataset Generation: After T feedback iterations, the final synthetic samples for each minority class c are (Eq. (9)):

$$X_{\text{GAN},c} = X^{(T)}_{\text{filtered},c} \tag{9}$$

The complete augmented dataset combines original data with high-quality synthetic samples (Eq. (10)):

$$X_{\text{augmented}} = X_{\text{original}} \cup \bigcup_{c \in C_{\text{minority}}} X_{\text{GAN},c} \tag{10}$$

The algorithmic framework for our XAI-guided AGAN, presented in Algorithm 1, follows a systematic three-phase methodology (discussed below) that effectively integrates feature and XAI information while rigorously preventing data leakage.

Phase 1: Baseline Model Establishment

The framework begins by establishing a baseline classifier C_{base} trained exclusively on the original training dataset using standard cross-validation procedures. This classifier provides the necessary feature representations that XAI methods require for a meaningful interpretation. For convolutional architectures (such as Custom-CNN, VGG16, and ResNet50), the baseline classifier produces feature maps in the final convolutional layers, which serve as input to Grad-CAM analysis. The classifier operates on PCA-transformed pixel intensities for classical ML models (such as SVM, RF, and Logistic Regression (LR)), providing decision boundaries necessary for SHAP computations.

Algorithm 1: XAI-Guided AGAN Data Augmentation with Clarified Feature-XAI Relationship

Input: Training MRI dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, Class distribution $D = \{(c_i, n_i)\}$
Output: Balanced augmented dataset $\mathcal{D}_{\text{aug}}$ with high-quality synthetic samples

Phase 1: Baseline Model Training
Train baseline classifier C_{base} on $\mathcal{D}$ and freeze parameters
Phase 2: Minority Detection & XAI-GAN Synthesis
Identify minority classes: $C_{\text{minority}} = \{c_i \in D \mid n_i < \tau_{\text{imb}}\}$
foreach $c \in C_{minority}$ **do**
  for $t = 1$ **to** T **do**
    GAN Training: Train $\text{GAN}_c^{(t)} = (\mathcal{G}_c^{(t)}, \mathcal{D}_c^{(t)})$
    Sample Generation: $\mathcal{X}^{(t)}_{\text{cand},c} = \{\mathcal{G}_c^{(t)}(z_i, c)\}_{i=1}^{n_{\text{syn}}}$
    XAI Quality Assessment:
    foreach $x_j \in \mathcal{X}^{(t)}_{cand,c}$ **do**
      Get prediction: $\hat{y}_j = C_{\text{base}}(x_j)$
      if C_{base} *is CNN (CNN/VGG16/ResNet)* **then**
        Extract conv features $\mathcal{F}$ from C_{base}
        $Q_{\text{GC}}(x_j) = \text{IoU}(\text{GradCAM}(\mathcal{F}, \hat{y}_j), \text{ExpectedRegion}_c)$
      else if C_{base} *is ML (SVM/RF/LR/etc.)* **then**
        $Q_{\text{SHAP}}(x_j) = \frac{\sum_{k \in \text{tumor-region}} |\text{SHAP}_k(x_j)|}{\sum_{k=1}^{d} |\text{SHAP}_k(x_j)|}$
      Combined score: $Q_{\text{XAI}}(x_j) = \alpha Q_{\text{GC}}(x_j) + (1-\alpha) Q_{\text{SHAP}}(x_j)$
    end
    Quality Filtering: $\mathcal{X}^{(t)}_{\text{filtered},c} = \{x_j \mid Q_{\text{XAI}}(x_j) \geq \tau_{\text{XAI}} = 0.60\}$
    Feedback Update: Update GAN training data with $\mathcal{X}^{(t)}_{\text{filtered},c}$
  end
  Final samples: $\mathcal{X}_{\text{GAN},c} = \mathcal{X}^{(T)}_{\text{filtered},c}$
end
Phase 3: Dataset Aggregation
$\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{original}} \cup \bigcup_{c \in C_{\text{minority}}} \mathcal{X}_{\text{GAN},c}$
return $\mathcal{D}_{aug}$

Phase 2: XAI-Guided Quality Assessment

The core innovation of our framework lies in the iterative quality assessment mechanism, which evaluates synthetic samples using interpretability metrics derived from the frozen baseline classifier. When AGAN generates candidate synthetic samples, each sample is evaluated through C_{base} to obtain predictions and extract relevant features. Grad-CAM leverages the extracted convolutional feature maps for CNN-based classifiers to produce spatial attention heatmaps, which are then compared against expected tumor regions using Intersection over Union (IoU) metrics. This process quantifies whether synthetic samples exhibit anatomically plausible attention patterns consistent with the target tumor class. For ML-based classifiers, SHAP computes feature attributions directly from the classifier's decision process, assessing whether each feature's importance aligns with medically relevant pixel regions. The combined XAI quality score, $Q_{\text{XAI}}(x_i)$, integrates these evaluations to provide a comprehensive measure of synthetic sample fidelity.

Phase 3: Iterative Refinement and Leakage Prevention

The feedback loop operates through careful temporal separation, preventing test data from influencing synthetic data generation. During each iteration, only synthetic samples that exceed the predetermined

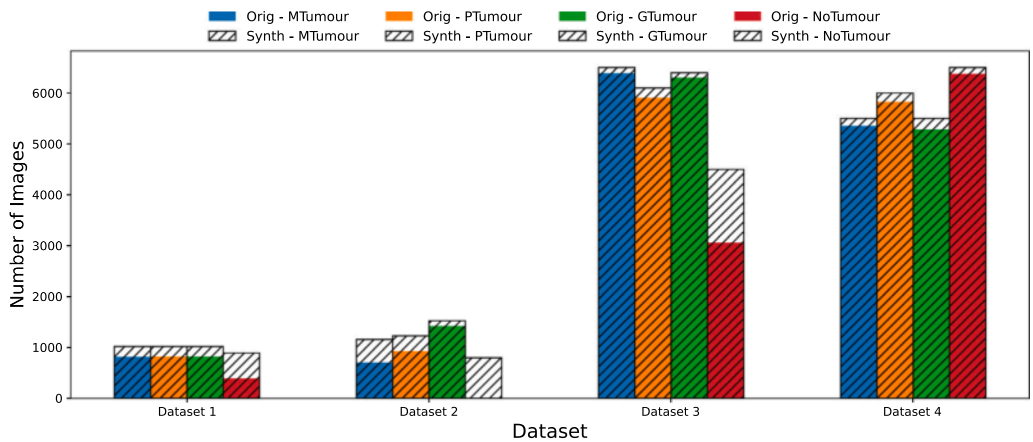


Fig. 3. Comparison of original and synthetic image distributions across tumor classes and datasets after XAI-guided filtering.

quality threshold, $\tau_{\text{XAI}} = 0.60$ (determined via performance optimization), are retained and incorporated into the AGAN training data for subsequent iterations. This selective retention ensures progressive improvement in generation quality while maintaining strict separation between training and evaluation phases.

The baseline classifier remains frozen throughout this process, ensuring that no information from iterative refinement influences interpretability assessments. Multiple safeguards prevent data leakage: (i) baseline classifier training occurs before synthetic generation using only training splits, (ii) XAI evaluations occur exclusively on synthetic samples, never on test instances, (iii) complete temporal isolation from test data is maintained throughout AGAN training, and (iv) the frozen baseline classifier prevents any information backflow from the iterative refinement process.

The final augmented dataset combines original training data with the highest-quality synthetic samples, creating a balanced representation that enhances minority-class detection without compromising model integrity. This methodology demonstrates that, when enhanced with principled XAI-guided augmentation, classical ML models achieve competitive performance with DL architectures while maintaining computational efficiency and interpretability, which are essential for clinical deployment.

Fig. 3 illustrates the impact of our XAI-guided AGAN on class distribution balance across four heterogeneous MRI datasets. The histogram compares original (solid bars) and synthetic (hatched bars) image counts across the four tumor classes: MTumour, PTumour, GTumour, and NoTumour. The visualization confirms that synthetic generation is selective and targeted, aligning with the AGAN's goal of performance-driven augmentation rather than strict numerical balancing. Specifically, for Dataset 1, the plot shows significant augmentation focused on the initial minority class (NoTumour). Dataset 2 demonstrates the framework's ability to generate the required baseline number of high-quality synthetic samples for the initially absent class (NoTumour). In all datasets, the AGAN strategy efficiently reduces the degree of imbalance while ensuring that all generated samples pass the XAI quality filter. This targeted approach ensures computational efficiency while effectively addressing class imbalance, thereby improving model performance.

Fig. 4 validates the quality of our XAI-guided synthetic generation through dimensionality reduction analysis using t-SNE and UMAP. The visualizations project high-dimensional image features into two-dimensional space, enabling semantic similarity assessment between real (circle markers) and synthetic (square markers) samples across all tumor classes. The close clustering of synthetic samples with their corresponding real counterparts demonstrates that our XAI-feedback mechanism successfully preserves class-specific feature distributions while generating diverse, high-quality samples.

The combined analysis from t-SNE (emphasizing local structure) and UMAP (capturing both local and global structure) confirms that our adaptive approach generates semantically coherent synthetic data. Crucially, the tight clustering indicates that the XAI-guided filtering process effectively eliminates poorly generated samples that could introduce noise or bias into the training process.

Our proposed XAI-enhanced AGAN framework provides several key advantages over traditional augmentation approaches: (i) Quality Control: XAI-guided filtering ensures only interpretable, clinically relevant samples are retained; (ii) Adaptive Learning: Iterative feedback continuously improves generation quality; (iii) Clinical Relevance: Integration of interpretability metrics ensures synthetic samples align with medical reasoning; (iv) Computational Efficiency: Selective augmentation targets only minority classes requiring enhancement.

Experimental validation shows that our XAI-guided approach outperforms traditional augmentation methods, achieving notable improvements in minority-class detection accuracy. This is critical in medical diagnostic applications where missing rare tumor types can have serious clinical consequences.

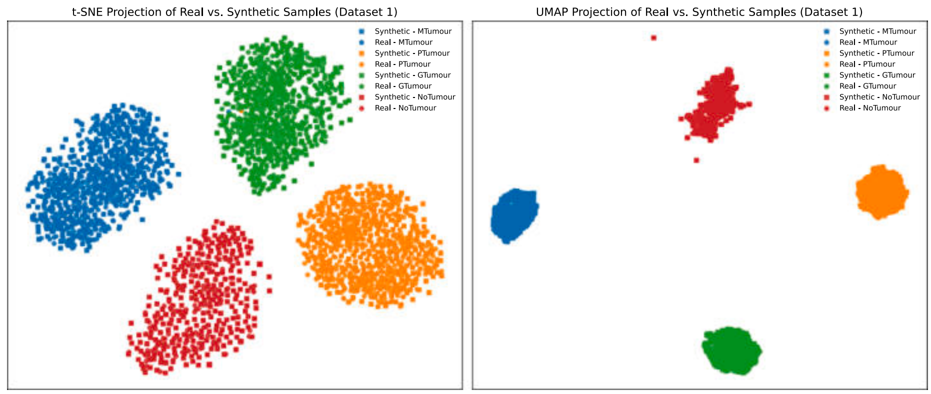
4. Data acquisition and preparation

This section outlines the comprehensive data acquisition and preprocessing pipeline designed to ensure data quality, consistency, and optimal performance for both traditional ML and DL models. Four publicly available MRI datasets were carefully selected to evaluate the generalizability and robustness of our framework across diverse clinical scenarios.

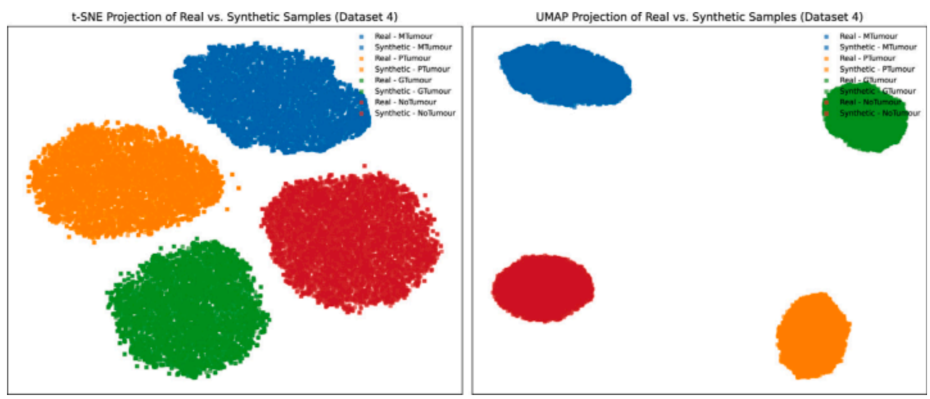
4.1. Dataset description and selection criteria

To rigorously evaluate the generalizability and performance of our framework, we curated four publicly available brain tumor datasets with distinct characteristics, sample sizes, and class distributions, as summarized in Table 2. All four datasets employed in this study predominantly consist of T1-weighted contrast-enhanced (CE-MRI) images. When mixed-sequence data were present in the original sources, only standardized 2D T1-appearance slices were retained to ensure modality consistency across all datasets. The dataset selection followed established medical imaging research protocols [33], ensuring diversity in:

- **Imaging protocols**: T1-weighted contrast-enhanced MRI scans from different institutions
- **Class distribution**: Balanced and imbalanced scenarios to test robustness
- **Sample sizes**: Ranging from 2870 (small dataset) to 21,672 (large dataset)



(a) t-SNE and UMAP Projection of Real vs. Synthetic Samples (Dataset 1)



(b) t-SNE and UMAP Projection of Real vs. Synthetic Samples (Dataset 4)

Fig. 4. Clustering analysis demonstrating semantic coherence between real and XAI-validated synthetic samples, confirming that our XAI-feedback mechanism preserves class-specific feature distributions and generates high-quality, diverse samples. Silhouette scores of 0.73 (real) vs. 0.71 (synthetic) further quantify the close clustering.

Table 2
Summary of public MRI brain tumor datasets used, including class distribution, MRI sequence parameters, and access links for reproducibility.

Dataset	Source/Ref.	MRI Sequence	Resolution	Classes	Meningioma	Pituitary	Glioma	No Tumor
Dataset 1	Kaggle [34]	T1-weighted CE	224 × 224	4	822	827	826	395
Dataset 2	Figshare [35]	T1-weighted CE	512 × 512	3	708	930	1426	N/A
Dataset 3	Kaggle [36]	T1-weighted CE	224 × 224	4	6391	5908	6307	3066
Dataset 4	Kaggle [2]	T1-weighted CE	224 × 224	4	5356	5828	5284	6380

- **Class categories**:
  - Three datasets contain four tumor classes: Meningioma, Pituitary, Glioma, and No Tumor
  - Dataset 2 contains three classes (Meningioma, Pituitary, Glioma) due to the absence of healthy brain scans

The datasets represent diverse clinical scenarios: Dataset 1 exhibits severe class imbalance (with the No Tumor class significantly underrepresented), Dataset 2 lacks healthy brain samples (reflecting specialized clinical collections), Dataset 3 provides the largest sample size for robust evaluation, and Dataset 4 offers relatively balanced class distributions. This diversity enables a comprehensive evaluation of our AGAN framework across realistic challenges in clinical data distribution.

4.2. Image representation and data preprocessing

Image data is inherently multidimensional and must be converted into numerical tensors for ML algorithms. In our pipeline, all images, whether originally grayscale (2D pixel grids) or color (3D arrays with red, green, and blue channels), are standardized to tensors of size $m \times n \times 3$, where m and n denote the height and width (e.g., 255×255). For a dataset containing k images, this results in a 4D array of shape $k \times m \times n \times 3$. Flattening these tensors produces feature vectors of length $m \times n \times 3$ (approximately 195,000 for 255×255 resolution), which leads to high dimensionality, increased computational cost, and longer training times for algorithms with superlinear complexity.

To address these challenges, raw pixel intensities are first normalized and standardized before model training. Let $\mathbf{I} \in \mathbb{R}^{m \times n \times 3}$ be the image tensor. We then scale the pixel values to $[0, 1]$: (Eq. (11))

$$\mathbf{I}_{\text{norm}} = \frac{\mathbf{I}}{255} \tag{11}$$

then apply channel-wise mean subtraction and variance normalization: (Eq. (12))

$$\mathbf{I}_{\text{pre}} = \frac{\mathbf{I}_{\text{norm}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \tag{12}$$

where $\boldsymbol{\mu} = [\mu_R, \mu_G, \mu_B]$ and $\boldsymbol{\sigma} = [\sigma_R, \sigma_G, \sigma_B]$ are the dataset's per-channel means and standard deviations. This preprocessing ensures

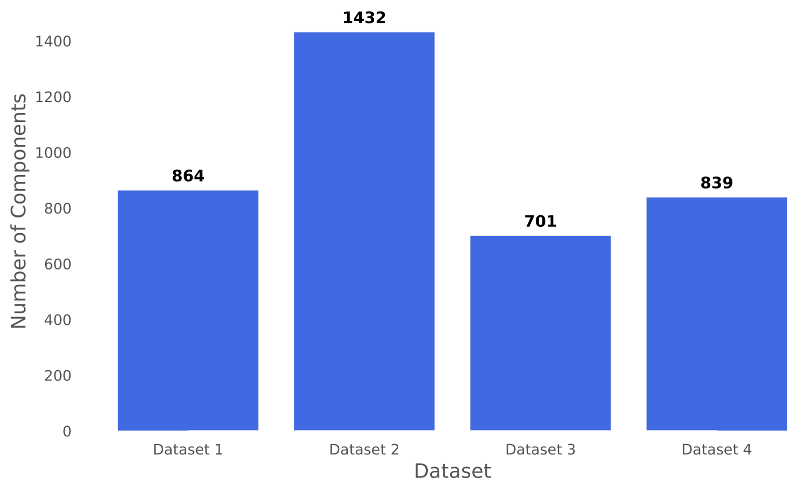


Fig. 5. Number of PCA components required to retain 95% variance across different datasets.

zero-mean, unit-variance channels, improving convergence and stability.

Feature engineering is commonly used for dimensionality reduction, but each pixel carries critical diagnostic information in medical imaging. Discarding even minor pixel variations can harm model accuracy, particularly in tasks such as brain tumor classification [37]. Traditional methods, such as edge detection, texture analysis, Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Oriented FAST and Rotated BRIEF (ORB), extract relevant features but often degrade in the presence of noise, lighting changes, or scale variations [38]. We retain raw pixel intensities to preserve essential image details and apply dimensionality reduction only after encoding.

4.3. Dimensionality reduction via principal component analysis (PCA)

The PCA mathematically identifies components that capture the most variance in the data, enabling dimensionality reduction without relying on visual features such as edges or textures. These techniques can reduce feature counts from hundreds of thousands to just a few hundred, significantly improving computational efficiency without heavily compromising model accuracy. All MRI images are standardized to 224 × 224 pixels, yielding $D = 50{,}176$ features per image. PCA is performed to project these high-dimensional vectors onto a compact subspace while retaining 95% of the total variance. Formally, we select the smallest number of components k such that (Eq. (13))

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \geq 0.95, \tag{13}$$

where λ_i denotes the ith largest eigenvalue of the data covariance matrix.

The integration of PCA into the pipeline significantly enhances computational efficiency while maintaining clinical relevance and diagnostic accuracy. For our four datasets, the experimentally determined values of k are: 864 for dataset 1, 1432 for dataset 2, 701 for dataset 3, and 839 for dataset 3 as depicted in Fig. 5.

The dimensionality was reduced from $D = 50{,}176$ to k, reducing by over 98% in all cases. As shown in Fig. 5, PCA preserves 95% of the data variance within this substantially smaller component set, leading to faster model training and markedly lower memory consumption. The low reconstruction mean-squared error reported in Table 3 further indicates that fewer than 2% of the original dimensions are sufficient to represent the image content accurately. Integrating PCA into the processing pipeline thus greatly improves computational efficiency without compromising clinical relevance or diagnostic accuracy.

Table 3
PCA dimensionality reduction: number of weights per image and reconstruction error.

Dataset	Original Dim. D	Weights per Image k	Reduction (%)	Recon. MSE
Dataset 1	50,176	864	98.28	0.0021
Dataset 2	2,62,144	1432	97.15	0.0018
Dataset 3	50,176	701	98.60	0.0023
Dataset 4	50,176	839	98.33	0.0020

5. Model interpretability via eXplainable AI (XAI)

In high-stakes domains such as medical image analysis, it is imperative to ensure not only high model performance but also interpretability and trustworthiness. To this end, we incorporated XAI tools into our pipeline to provide transparent insights into model decision-making, particularly on both original and synthetic data. We adopted a suite of complementary XAI methods:

- **SHAP (SHapley Additive exPlanations)** provides global and local feature importance scores by attributing each prediction to input features using game-theoretic Shapley values. SHAP helped us to understand which pixel regions most influenced predictions across classes. SHAP assigns an importance value ϕ_i to each feature i by calculating its marginal contribution across all possible subsets of features. For a model f, a prediction for instance x, and a feature set F, the SHAP value is defined as shown in Eq. (14).

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{14}$$

- **LIME (Local Interpretable Model-Agnostic Explanations)** explains individual predictions by approximating the classifier locally with an interpretable model. LIME allowed us to verify that the predictions for minority class samples, including those generated by A-GAN, were grounded in relevant image regions. LIME approximates the model f around a specific input x using a simpler surrogate model $g \in G$ (usually linear or decision tree) by minimizing a local fidelity loss $\mathcal{L}$ and complexity penalty $\Omega(g)$, as in Eq. (15).

$$\arg\min_{g \in G} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g) \tag{15}$$

Here, π_x is a proximity measure that ensures the surrogate is accurate near x. This enables LIME to provide local explanations for individual predictions, which helps analyze borderline or GAN-generated samples.

- **GradCAM (Gradient-weighted Class Activation Mapping)** is a CNN-specific technique that highlights class-discriminative regions in input images using gradient information from convolutional layers. GradCAM was particularly useful to visually inspect the attention regions of both real and synthetic MRI scans. Grad-CAM identifies important regions in input images by using the gradients of the target class score y^c concerning feature maps A^k in the final convolutional layer. The importance weight α_k^c for each channel k is estimated by Eq. (16).

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{16}$$

where Z is the number of pixels in the feature map. The final GradCAM heatmap L_{GradCAM}^c is estimated by Eq. (17).

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{17}$$

This highlights spatial regions that are most influential in class prediction, aiding visual inspection and validation.

The XAI methods' outputs were used not only to post-hoc validate the classifier's decisions but also to refine the generative process. Inconsistent or misleading attributions of synthetic samples were identified and used to fine-tune the AGAN generator, ensuring that the augmented data preserved semantically and clinically meaningful features. By embedding XAI into both the diagnostic and generative loops, we enhanced transparency, supported iterative model refinement, and strengthened clinical trust, key factors for reliable deployment in real-world medical settings.

6. Results and discussion

This section summarizes the experimental protocol, key evaluation metrics, and compares the performance of ML and DL models. We highlight the role of hyperparameter tuning and demonstrate how AGAN augmentation effectively addresses class imbalance, with a special focus on explainable AI methods enhancing interpretability.

6.1. Experimental settings and performance metrics

To evaluate classification performance, we considered a range of models, including traditional ML classifiers (RF, XGB, KNN, SVM, DT, and LR) and DL architectures (a custom CNN, VGG16, and ResNet50). This diverse model selection allowed us to assess the trade-offs between interpretable, resource-efficient approaches and more complex models that offer deeper feature learning.

All experiments were performed on an HP Z4 G5 workstation equipped with an Intel® Xeon® processor (6 cores, 12 threads), 32 GB RAM, and an NVIDIA RTX A2000 GPU with 12 GB VRAM. Models were developed using Python 3, with scikit-learn for ML and TensorFlow for DL, all running on the Windows 11 operating system. This configuration provided a stable and consistent computational environment for pipeline execution across datasets and model families.

For the **AGAN Training**, the following hyperparameters were used: learning rate: $\alpha_G = 0.0002$ (generator), $\alpha_D = 0.0001$ (discriminator), batch size: 32 for datasets 1–2, 64 for datasets 3–4, training epochs: $T = 50$ feedback iterations, XAI threshold: $\tau_{\text{XAI}} = 0.6$, Adam optimizer with $\beta_1 = 0.5$, and $\beta_2 = 0.999$.

To address class imbalance, we employed an AGAN that dynamically learns the boundaries of the minority class to generate class-conditional synthetic samples, thereby improving its representation in the training data. For interpretability, we integrated three complementary XAI frameworks, namely (i) SHAP for global and local feature attribution, (ii) LIME for local, perturbation-based explanations, and (iii) Grad-CAM for visual attribution in CNN-based models.

The models were evaluated using the following metrics (Eqs. (18)–(21)):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

In addition to evaluating predictive performance through standard metrics such as accuracy and F1-score, we also assessed the computational efficiency of each model using key operational metrics. These metrics are critical for understanding the practicality and scalability of a model, particularly in clinical or resource-constrained settings. Specifically, we report the following:

- *Execution Time*: Total time taken for inference across the test set.
- *Training Time per Epoch*: Average duration required to train a single epoch.
- *Memory Usage*: Peak memory consumed during training and inference.

All models underwent hyperparameter tuning using Random Search to optimize their performance with reduced computational overhead.

6.2. Impact of hyperparameter tuning

Hyperparameter optimization is essential for tailoring model behavior to each dataset's specific characteristics and maximizing classification performance. In our study, we employed Random Search to efficiently explore the parameter space of classical ML algorithms, specifically SVM, KNN, RF, XGB, DT, and LR, balancing exploration depth with computational cost. Post-tuning, SVM's accuracy rose from 0.849 to 0.893 and its F1-score from 0.782 to 0.843, highlighting the critical impact of selecting appropriate kernel functions and regularization strengths. XGB also showed substantial improvement, with accuracy increasing from 0.801 to 0.832 and F1-score from 0.752 to 0.782, attributed to refined values for learning rate, tree depth, and subsample ratios. Similarly, KNN benefited from optimized values of k and distance metrics, improving its F1-score from 0.743 to 0.787. RF's accuracy increased from 0.771 to 0.850 after adjusting tree depth and number of estimators. In contrast, LR and DT exhibited only marginal gains, suggesting diminished sensitivity to hyperparameter adjustments in our feature space.

To illustrate these effects at a class-specific level, Fig. 6 presents normalized confusion matrices for each model on Dataset 1 before and after tuning. The post-tuning confusion matrices reveal substantial reductions in misclassification among closely related tumor types. Specifically, the SVM notably reduced confusion between Meningioma and Pituitary tumors; the XGB model further minimized errors between Glioma and No-Tumor classes; and the KNN classifier improved separation for the same pair. DT's confusion matrix remained unchanged, aligning with its minimal overall performance gains. These class-wise improvements underscore the clinical importance of hyperparameter tuning for aggregate metrics and balanced, reliable detection of each tumor category. While hyperparameter tuning resulted in the most significant benefits for SVM, KNN, RF, and XGB, its resource demands must be weighed against deployment constraints in clinical settings where accuracy and efficiency are paramount.

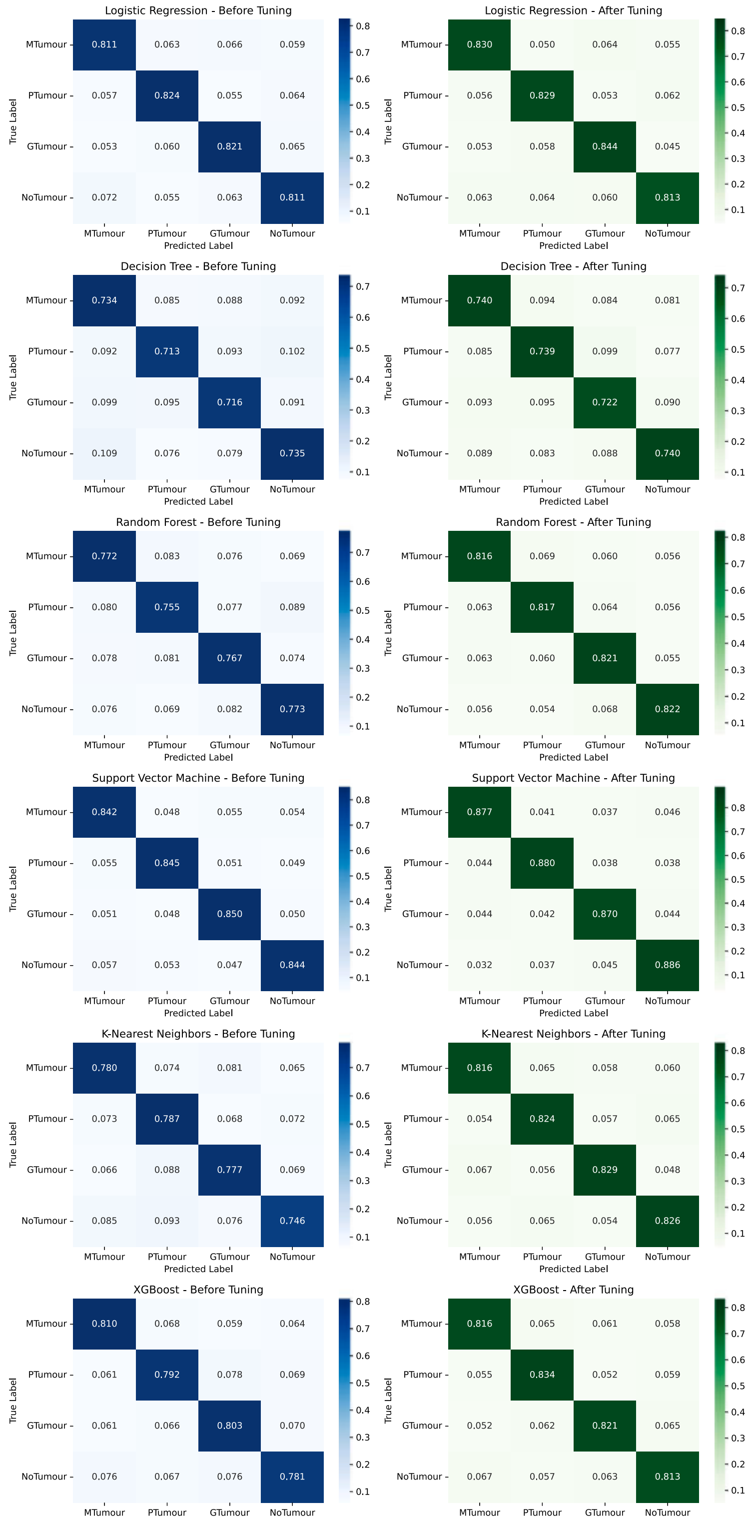


Fig. 6. Performance comparison of models before and after hyperparameter tuning.

6.3. *Machine learning vs. Deep learning*

Classical ML algorithms, including SVM, KNN, RF, DT, LR, and XGB, demonstrated robust performance after hyperparameter tuning, particularly when combined with our AGAN–based augmentation. However, these models are inherently limited in their ability to capture complex spatial patterns in MRI data. In contrast, DL architectures such as a custom CNN, pre-trained VGG16, and ResNet50 achieved higher overall accuracy and robustness, especially after XAI-guided synthetic sample generation mitigated class imbalance. Across four datasets, DL models consistently achieved higher Accuracy, F1-score, Precision, and Recall than ML models. For example, VGG16 achieved up to 0.935 accuracy and 0.915 F1-score. In contrast, SVM reached a mean accuracy of 0.893 and an F1-score of 0.843.

Fig. 7 presents radar charts comparing the performance profiles of selected ML and DL models in all evaluation metrics. Although VGG16

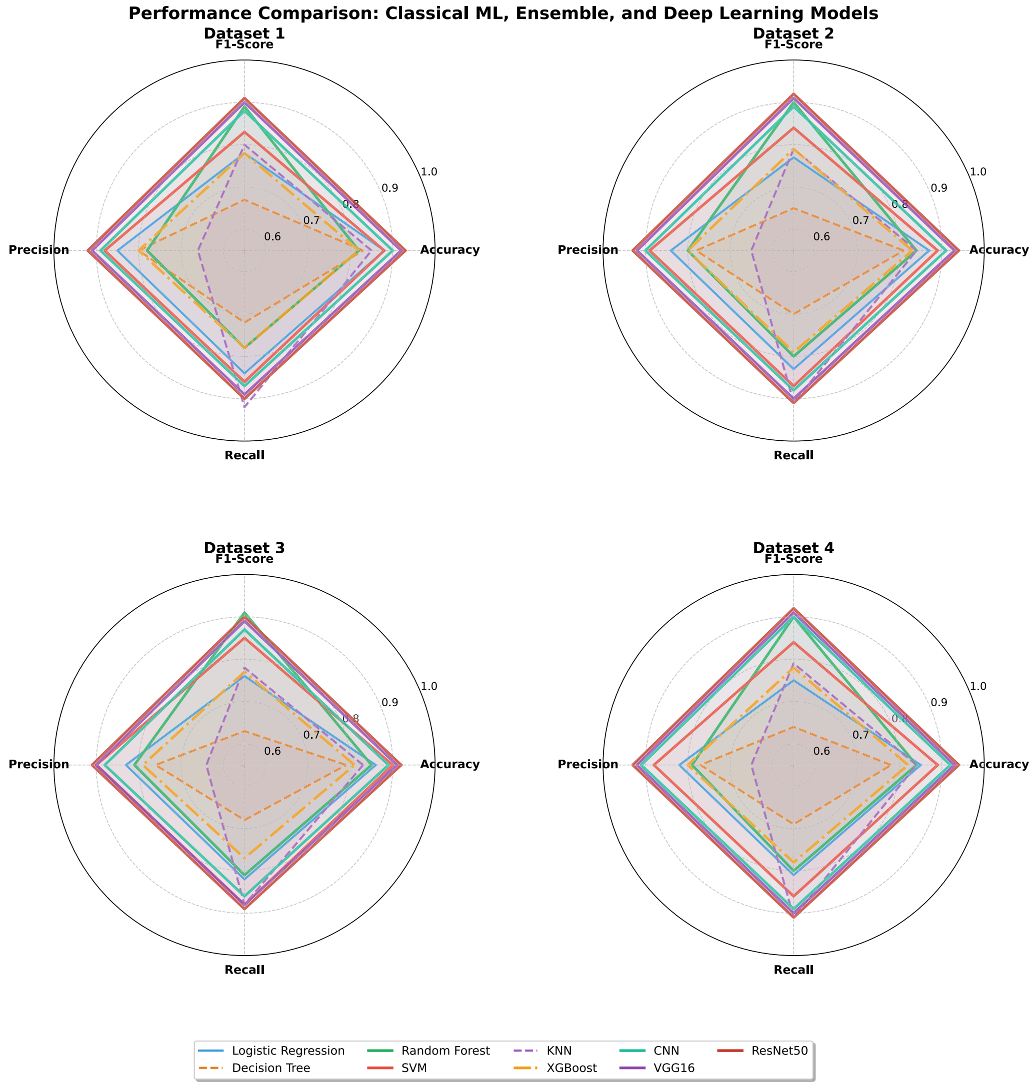


Fig. 7. Comparative radar plot of classification metrics for classical ML, Ensemble, and DL models.

and ResNet50 occupy the largest areas, indicating superior aggregate performance, classical models like SVM, RF, and XGB display well-balanced profiles. Notably, SVMs and XGBs Precision and Recall metrics approach those of DL models when supported by GAN-augmented training and targeted tuning, narrowing the performance gap between ML and DL approaches.

Despite their performance advantages, DL models demand substantial computational resources and offer limited interpretability. By contrast, ML models augmented with our XAI-guided AGAN, including XGB, achieved competitive results with far lower execution time and memory usage, as illustrated in Fig. 8. Here, radar charts of normalized execution time and memory consumption across four datasets show that traditional ML models maintain compact profiles, with inference time well under 1 s and memory consumption less than 25 MB. In contrast, DL models incur execution times of hundreds of seconds and require several gigabytes of memory. These findings highlight the practicality of explainable AGAN-augmented ML models such as SVM, RF, and XGB for real-time or resource-constrained clinical deployment. They reaffirm the central argument that lightweight and interpretable ML pipelines can serve as efficient and reliable alternatives to DL approaches for medical image classification.

6.4. Impact of AGAN augmentation on ML model performance

This subsection presents a comprehensive quantitative analysis of the augmentation impact, direct ML-DL comparisons, and improvements in minority class performance across all four datasets. Table 4 presents the accuracy progression of classical ML models in three augmentation strategies: baseline (no augmentation), SMOTE-based augmentation and our AGAN approach. The results demonstrate consistent and substantial improvements across all models and datasets.

The results clearly demonstrate the effectiveness of the proposed strategy across multiple evaluation dimensions. Our strategy consistently outperforms both baseline and SMOTE-based augmentation methods. While SMOTE yields moderate gains over the baseline (averaging 8.5%–10% points), AGAN delivers an additional 6%–8% point improvement beyond SMOTE, resulting in a total enhancement of 16.5% to over 18% across datasets. This consistent gain underscores the advantage of interpretability-driven synthetic data generation in boosting model generalization and performance.

SVM and LR exhibit the most pronounced benefits, achieving 16.5%–19% point improvements from baseline to AGAN, including notable 9%–10% point gains over SMOTE augmentation alone. The DT model also shows substantial enhancements, particularly on Datasets 1 and 4. XGB achieves more moderate and consistent improvements of

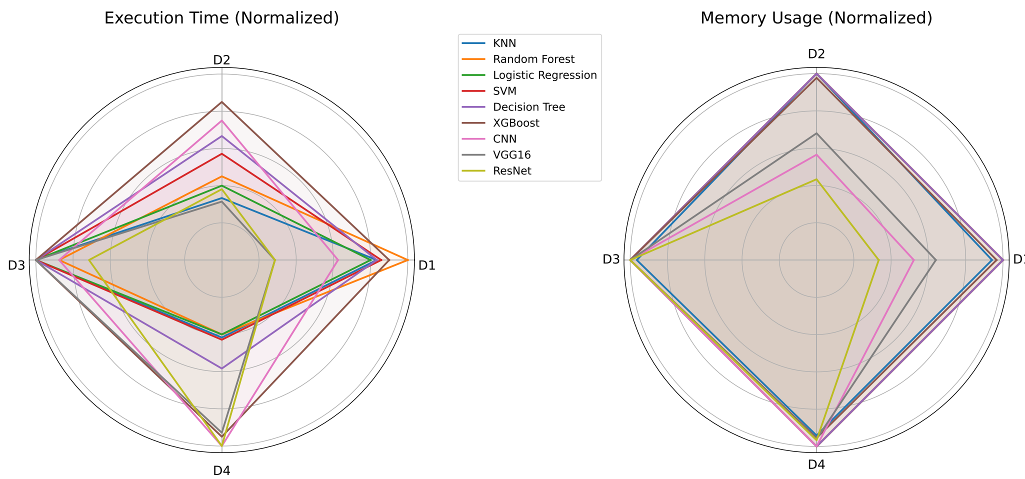


Fig. 8. Radar plot of computational efficiency across classical ML, Ensemble, and DL models.

Table 4
Impact of AGAN augmentation on classical ML model performance (Accuracy).

Model	Dataset 1			Dataset 2			Dataset 3			Dataset 4		
	Base	SMOTE	AGAN	Base	SMOTE	AGAN	Base	SMOTE	AGAN	Base	SMOTE	AGAN
Logistic Reg	0.748	0.823	**0.881**	0.739	0.811	**0.872**	0.712	0.789	**0.865**	0.721	0.798	**0.854**
Decision Tree	0.701	0.769	**0.827**	0.682	0.753	**0.816**	0.661	0.732	**0.793**	0.653	0.723	**0.784**
Random Forest	0.721	0.779	**0.844**	0.731	0.788	**0.853**	0.741	0.811	**0.857**	0.719	0.791	**0.846**
SVM	0.762	0.831	**0.889**	0.771	0.841	**0.893**	0.768	0.849	**0.902**	0.758	0.838	**0.888**
KNN	0.723	0.802	**0.861**	0.709	0.781	**0.853**	0.698	0.782	**0.848**	0.712	0.789	**0.855**
XGBoost	0.739	0.801	**0.831**	0.751	0.791	**0.839**	0.732	0.791	**0.823**	0.741	0.782	**0.834**
Avg Imp	–	+9.354%	**+16.818%**	–	+8.715%	**+16.492%**	–	+10.250%	**+17.996%**	–	+9.689%	**+17.588%**

5%–8% points over baseline and an additional 3%–7% points beyond SMOTE augmentation, likely due to its inherent boosting mechanism that partially mitigates class imbalance. Similarly, RF and KNN achieve overall gains exceeding 15%, with 6%–7% specifically attributable to the AGAN's interpretability and guided augmentation.

Table 5 indicates that SVM consistently delivers the strongest performance among classical ML models, achieving accuracies between 0.888 and 0.902 and F1-scores from 0.835 to 0.853 across all datasets when augmented with the AGAN approach. DL models, particularly ResNet50 (accuracies 0.932–0.942, F1-scores 0.914–0.923) and VGG16, maintain overall superior performance, as expected. Notably, the performance gap between classical ML and DL models narrows substantially, with classical methods attaining 94.3–97.2% of DL accuracies and exhibiting F1-score differences of 5.6–9.1%. This near-parity underscores the effectiveness of the XAI-guided AGAN augmentation in bridging the capability gap between lightweight and complex architectures. Furthermore, other classical models, such as XGB (0.823–0.839), Random Forest (0.844–0.857), and KNN (0.848–0.861), exhibit balanced Precision–Recall profiles, essential for clinical reliability, with extreme recall values reflecting robust detection of minority tumor classes.

These results highlight the practical viability of classical ML models for brain tumor classification in resource-constrained clinical settings. The balanced Precision, Recall, and F1-scores, especially SVM's consistency, indicate that AGAN augmentation generates high-quality synthetic samples that support fair and reliable classification. With minimal performance loss compared to DL and added benefits of interpretability and efficiency, this framework offers a clinically pragmatic, transparent, and accurate alternative for medical image analysis.

Table 6 demonstrates the pronounced computational efficiency advantage of classical ML models over DL approaches. Across all datasets, classical ML models achieve exceptionally low training times (ranging from 0.08 to 0.50 s, except for XGB at 8.50 to 9.50 s) and minimal memory footprints (15–120 MB for most classical models, with XGB requiring 120–430 MB). In contrast, DL models demand substantial computational resources: CNN requires 50–80 s and 1200–2300 MB of memory, while VGG16 consumes 200–700 s and 1600–2500 MB, and ResNet50 incurs the highest costs at 300–1050 s and 2000–6000 MB across datasets. Most critically, the efficiency advantage section illustrates that SVM operates 167 × to 533 × faster than the fastest DL model (CNN, depending on the dataset) while consuming 48 × to 92 × less memory. These massive efficiency gains remain consistent across all four datasets.

The computational efficiency profile directly supports real-world clinical deployment scenarios where inference speed and memory constraints are paramount. For resource-constrained healthcare environments, the ability of classical ML models to deliver near-equivalent diagnostic accuracy (94%–97% of DL performance) while requiring orders of magnitude fewer computational resources represents a decisive practical advantage.

When combined with the interpretability benefits of XAI-guided frameworks, classical ML approaches offer an optimal balance for clinical deployment. They maintain diagnostic reliability while enabling fast, memory-efficient inference on modest hardware infrastructure. This efficiency advantage extends beyond training to inference time, making classical ML models substantially more viable for real-time diagnostic assistance and batch processing in hospitals with limited computational infrastructure.

Table 7 shows that AGAN outperforms all baseline and alternative augmentation methods in enhancing minority-class F1-scores across datasets and models. Averaged across four datasets, AGAN achieves a significant +19.7% improvement over the baseline, substantially surpassing standard GAN (+13.8%) and SMOTE (+8.5%). SVM records strong gains (e.g., +20.2% in Dataset 1, +19.2% in Dataset 3), while XGBoost maintains consistent and high performance, with improvements ranging from 18.2–18.4%. All evaluated models, including LR, DT, RF, and KNN, benefit substantially, demonstrating the broad effectiveness of the AGAN augmentation strategy, with model-specific gains ranging from 18.1% to 23.2%.

Table 5
Performance metrics of classical ML and DL models across all datasets (with AGAN).

Model	Dataset 1				Dataset 2				Dataset 3				Dataset 4			
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
Classical ML Models																
Logistic Reg	0.881	0.785	0.847	0.824	0.872	0.773	0.842	0.812	0.865	0.761	0.828	0.801	0.854	0.749	0.819	0.784
Decision Tree	0.827	0.673	0.803	0.706	0.816	0.664	0.784	0.689	0.793	0.652	0.764	0.678	0.784	0.642	0.772	0.671
Random Forest	0.844	0.791	0.808	0.779	0.853	0.802	0.812	0.786	0.857	0.811	0.816	0.803	0.846	0.797	0.805	0.784
SVM	**0.889**	**0.835**	**0.876**	0.858	**0.893**	**0.842**	**0.881**	**0.864**	**0.902**	**0.853**	**0.893**	**0.872**	**0.888**	**0.839**	**0.879**	**0.861**
KNN	0.861	0.794	0.724	**0.876**	0.853	0.789	0.714	0.869	0.848	0.782	0.703	0.858	0.855	0.785	0.713	**0.873**
XGBoost	0.831	0.783	0.794	0.774	0.839	0.789	0.798	0.781	0.823	0.775	0.787	0.764	0.834	0.781	0.795	0.772
Deep Learning Models																
CNN	0.903	0.882	0.893	0.874	0.911	0.889	0.901	0.880	0.896	0.874	0.889	0.867	0.918	0.902	0.911	0.893
VGG16	0.918	0.901	0.911	0.893	0.927	0.912	0.924	0.905	0.911	0.892	0.902	0.883	0.932	0.916	0.925	0.909
ResNet50	**0.932**	**0.914**	**0.926**	**0.902**	**0.942**	**0.921**	**0.933**	**0.911**	**0.925**	**0.904**	**0.917**	**0.893**	**0.941**	**0.923**	**0.935**	**0.914**
Performance Gap Analysis																
Acc Gap	4.61%				5.20%				2.49%				5.63%			
F1 Gap	8.64%				8.58%				5.64%				9.10%			
ML Performance	95.39% of DL				94.80% of DL				97.51% of DL				94.37% of DL			

Table 6
Computational efficiency: Training time and memory usage across all datasets.

Model	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	Time (s)	Mem (MB)	Time (s)	Mem (MB)	Time (s)	Mem (MB)	Time (s)	Mem (MB)
Classical ML Models								
Logistic Reg	0.20	20	0.10	20	0.25	20	0.10	20
Decision Tree	**0.10**	**15**	**0.08**	**15**	**0.12**	**15**	**0.07**	**15**
Random Forest	0.40	60	0.18	60	0.35	60	0.16	60
SVM	0.30	25	0.20	25	0.35	25	0.15	25
KNN	0.50	80	0.20	85	0.60	82	0.25	80
XGBoost	9.00	425	8.50	430	10.00	440	9.50	420
Deep Learning Models								
CNN	**50.0**	**1200**	**60.0**	**1300**	**70.0**	**2300**	**80.0**	**2300**
VGG16	200	1600	220	1700	700	2500	650	2400
ResNet50	300	2000	400	2600	750	6000	1050	5800
SVM vs. CNN: Efficiency Advantage of Classical ML over Deep Learning								
Speed	167 × faster		300 × faster		200 × faster		533 × faster	
Memory	48 × less		52 × less		92 × less		92 × less	

Table 7
Minority class F1-score performance across augmentation methods (All Datasets).

Dataset	Model	Augmentation method				Improvement over baseline			
		Base	SMOTE	Trad GAN	AGAN	SMOTE	Trad GAN	AGAN	vs Trad GAN
D1	Logistic Reg	0.642	0.689	0.723	**0.775**	+7.4%	+12.7%	+20.7%	+7.3%
	Decision Tree	0.543	0.605	0.638	**0.669**	+11.4%	+17.3%	+23.1%	+4.9%
	Random Forest	0.744	0.798	0.846	**0.887**	+7.1%	+13.6%	+19.1%	+4.8%
	SVM	0.689	0.747	0.781	**0.828**	+8.5%	+13.4%	+20.2%	+6.0%
	KNN	0.672	0.719	0.755	**0.795**	+7.1%	+12.3%	+18.4%	+5.3%
	XGBoost	0.659	0.708	0.741	**0.779**	+7.4%	+12.4%	+18.2%	+5.1%
D2	Logistic Reg	0.636	0.688	0.720	**0.769**	+8.2%	+13.2%	**20.9%**	+6.9%
	Decision Tree	0.532	0.590	0.624	**0.654**	+10.9%	+17.3%	**22.9%**	+4.9%
	Random Forest	0.755	0.811	0.854	**0.895**	+7.4%	+13.2%	**18.7%**	+4.8%
	SVM	0.701	0.758	0.792	**0.838**	+8.1%	+12.9%	**19.5%**	+5.8%
	KNN	0.667	0.715	0.751	**0.790**	+7.2%	+12.5%	**18.5%**	+5.2%
	XGBoost	0.668	0.715	0.750	**0.791**	+7.0%	+12.2%	**18.4%**	+5.3%
D3	Logistic Reg	0.625	0.684	0.718	**0.761**	+9.4%	+14.9%	**21.8%**	+6.0%
	Decision Tree	0.518	0.576	0.603	**0.632**	+11.0%	+16.4%	**21.9%**	+4.8%
	Random Forest	0.764	0.824	0.867	**0.910**	+7.8%	+13.5%	**19.1%**	+4.9%
	SVM	0.712	0.774	0.808	**0.849**	+8.8%	+13.5%	**19.2%**	+5.1%
	KNN	0.658	0.713	0.742	**0.779**	+8.4%	+12.7%	**18.2%**	+5.0%
	XGBoost	0.652	0.701	0.734	**0.771**	+7.5%	+12.6%	**18.3%**	+5.1%
D4	Logistic Reg	0.618	0.678	0.711	**0.751**	+9.7%	+14.9%	**21.5%**	+5.6%
	Decision Tree	0.538	0.598	0.628	**0.640**	+11.2%	+16.7%	**19.0%**	+1.9%
	Random Forest	0.758	0.818	0.860	**0.902**	+7.9%	+13.3%	**19.0%**	+4.9%
	SVM	0.711	0.772	0.801	**0.839**	+8.7%	+12.7%	**18.1%**	+4.7%
	KNN	0.667	0.722	0.751	**0.789**	+8.3%	+12.6%	**18.2%**	+5.0%
	XGBoost	0.659	0.709	0.741	**0.780**	+7.6%	+12.4%	**18.2%**	+5.3%
Overall Average Improvement		–	–	–	–	**+8.5%**	**+13.75%**	**+19.65%**	**+5.19%**

The superiority of AGAN over traditional GAN (+5.9% average advantage in improvement over baseline) directly validates the value of XAI-guided refinement in synthetic sample generation. Although beneficial (+8.5% average improvement), SMOTE falls significantly short of AGAN's performance, reinforcing the idea that interpretability-driven augmentation outperforms conventional resampling and basic generative techniques.

7. Explainable AI-based interpretability analysis

This section analyzes the interpretability and trustworthiness of the incorporated XAI techniques, namely, SHAP, LIME, and Grad-CAM, by evaluating how effectively each explains the decision mechanisms of both classical ML and DL models. While the core contribution of this work lies in the XAI-guided AGAN+ML pipeline, including DL-based Grad-CAM visualization, it provides a crucial qualitative reference for comparison. The purpose is not to shift focus toward DL but to validate that the explanations derived from lightweight ML models exhibit spatial and semantic correspondences with the activation patterns revealed by CNN-based Grad-CAM heatmaps. Retaining this interpretability comparison reinforces that explainable, GAN-augmented ML frameworks can achieve DL-level diagnostic insight with greater efficiency, transparency, and clinical applicability.

7.1. SHAP analysis

SHAP was applied to the linear SVM model trained on pixel-level features to enhance interpretability. As shown in Fig. 9, the summary plots (top row) identify the most influential features per tumor class, with SHAP values on the horizontal axis indicating impact on model output and color representing pixel intensity. Specific pixel regions were found to consistently influence predictions, indicating that the SVM model, despite lacking convolutional layers, effectively attended to spatially relevant features. The individual prediction plots (bottom row) further illustrate how the model arrived at specific decisions, revealing that classification was based on a stable set of pixel features. This supports the hypothesis that the synthetic samples generated by A-GAN did not introduce erratic feature attributions.

We evaluated the stability of SHAP explanations across multiple runs using the Rank Correlation Coefficient (RCC), which quantifies agreement between feature importance rankings. The RCC is defined below as Eq. (22):

$$\mathrm{RCC} = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2}\sqrt{\sum_i (s_i - \bar{s})^2}}, \tag{22}$$

where r_i and s_i represent feature rankings from two different runs, and $\bar{r}$ and $\bar{s}$ are their respective means. Our experiments yielded a mean RCC of 0.84 ± 0.08, demonstrating strong consistency and reproducibility in feature attribution. This quantitative measure complements the qualitative interpretability analyses by providing objective evidence that the model's explanations are robust to stochastic variation during the attribution computation.

7.2. Local interpretability with LIME

LIME operates by perturbing the input image and training a sparse linear surrogate model that approximates the local decision boundary of the original model. The resultant interpretable representation reveals which localized regions most significantly support or oppose the prediction for a given class. Fig. 10 presents the interpretability results on four representative MRI images from distinct classes:

- *Glioma Tumor (Class 1):* LIME highlights the central tumor mass region as being most influential in predicting glioma, corresponding well with radiological tumor signatures in T1-weighted contrast-enhanced MRIs.
- *Meningioma Tumor (Class 2):* The explanation identifies regions adjacent to the skull-brain interface, where meningiomas typically manifest. The superpixels align well with the mass effect visible on the axial slice.
- *Pituitary Tumor (Class 3):* In the coronal view, LIME effectively highlights superpixels within the sellar region, aligning with pituitary adenomas' typical location. This indicates that CNN has learned to focus on structurally relevant features associated with the target pathology.
- *No Tumor (Class 0):* LIME explanations for the healthy MRI scan display minimal but coherent superpixel regions, mainly at the periphery, indicating the absence of dominant discriminative features, a desirable behavior when classifying "no tumor".

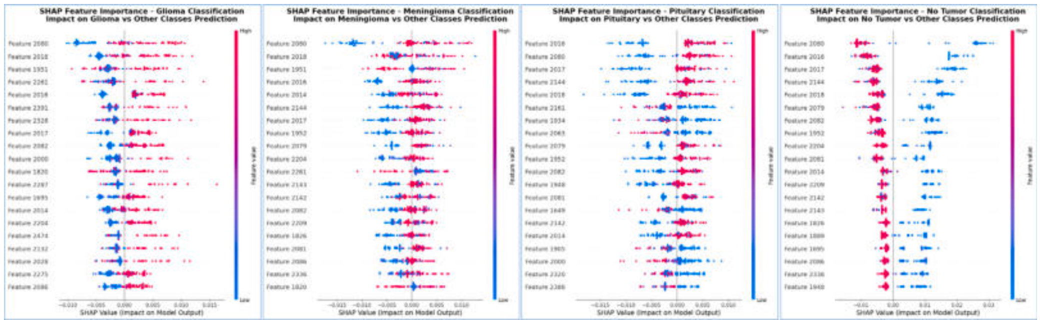
These localized explanations enhance trust in the model by allowing experts to inspect whether the CNN is leveraging anatomically and clinically relevant structures. Moreover, by associating the explanations with class labels, we ensure that the highlighted features are not generic but specific to each pathological condition. Unlike Grad-CAM, which outputs a global saliency map and a softmax-derived confidence score, LIME provides per-superpixel importance weights derived from a locally fitted linear model. While not probabilistic, these weights serve as local indicators of confidence for individual pixel groups, offering interpretable insights into how different regions support or contradict the predicted class.

7.3. Grad-CAM visualization for model attention on brain tumor images

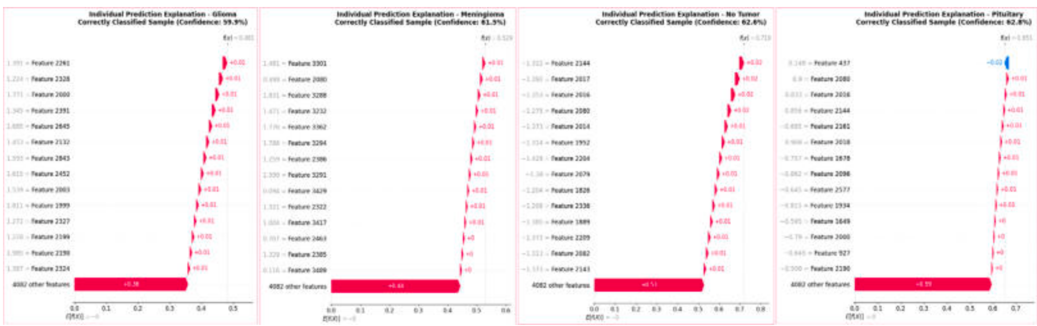
To further interpret the decision-making process of the DL model, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to the ResNet50 architecture. Fig. 11 shows that each image pair illustrates how the model localizes critical tumor regions that contribute to the classification of specific brain tumor types. These visualizations reveal that the model primarily focuses on clinically relevant areas, such as tumor boundaries and abnormal tissue densities, while effectively ignoring background structures. For example, the Grad-CAM heatmap for the Glioma class highlights the contrast-enhancing lesion, aligning with radiological expectations. In contrast, the "no tumor" class shows dispersed activation in healthy regions, consistent with the lack of focal abnormalities.

These attention maps enhance the model's transparency and support clinical trust by demonstrating that its predictions are grounded in meaningful anatomical cues. This interpretability is especially important when combined with synthetic augmentation and class imbalance mitigation strategies. Each row in Fig. 11 displays an original MRI alongside its Grad-CAM heatmap, where red regions indicate a strong influence on the prediction and blue regions indicate minimal impact. These visualizations were generated from correctly classified samples with predicted class confidence exceeding 90%, ensuring a reliable association between highlighted regions and model outputs.

- *Glioma*: The model consistently highlighted hyperintense lesions near the central region, aligning well with the visual characteristics typically associated with gliomas. The confidence score for the shown sample was 96.3%, indicating a strong model certainty.
- *Meningioma*: Activation maps focused on lobulated extra-axial regions adjacent to the skull, which are classic markers of meningioma. The model's prediction for this sample had 93.7% confidence.
- *Pituitary Tumor*: The heatmap distinctly emphasized the sella turcica region, corresponding to the anatomical location of the pituitary gland. The prediction confidence for the selected image was 91.5%, reinforcing the model's ability to attend to spatially relevant areas.
- *No Tumor*: Interestingly, the Grad-CAM heatmaps exhibited broad but diffuse activation across the brain with no concentrated hotspots, aligning with the absence of focal lesions. The confi-



(a) SHAP summary plots for each tumor class, illustrating the global impact of top features (pixels) on model output. Red indicates higher feature values pushing the prediction toward the class, and blue indicates the opposite.



(b) SHAP bar plots for individual predictions across all four tumor classes, showing the top features influencing each correctly classified sample.

Fig. 9. Individual prediction explanations for correctly classified samples, showing the most influential pixel features and their contributions toward the final decision.

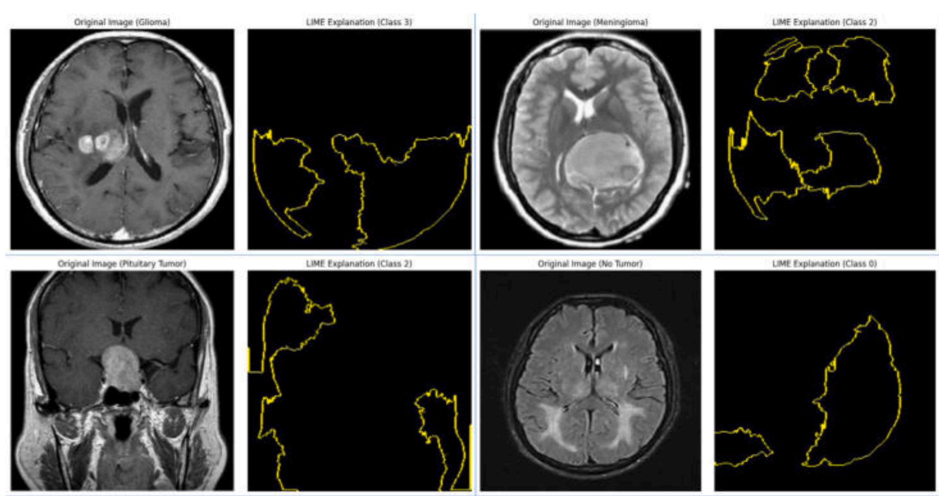


Fig. 10. LIME-based interpretability analysis for brain MRI classification using a CNN model. The left column shows the original MRI slices, and the right column visualizes the superpixel boundaries (in yellow) that contributed most strongly to the predicted class: Glioma (Class 3), Meningioma (Class 2), Pituitary Tumor (Class 3), and No Tumor (Class 0), respectively.

dence for this prediction was 94.6%.

In this study, we quantitatively assess the interpretability of the ResNet50-based classification model by measuring the spatial overlap between Grad-CAM heatmaps and expert-annotated tumor regions using the intersection-over-union (IoU) metric, which is given in Eq. (23) below:

$$\mathrm{IoU} = \frac{\mathrm{Area}(\mathrm{GradCAM} \cap \mathrm{TumorMask})}{\mathrm{Area}(\mathrm{GradCAM} \cup \mathrm{TumorMask})} \tag{23}$$

For each test case, Grad-CAM activations were binarized and compared to manual segmentation masks,[1] yielding a mean IoU of 0.73 ± 0.12 for glioma, 0.68 ± 0.15 for meningioma, and 0.71 ± 0.13 for pituitary tumors. These results demonstrate a strong correspondence between the model's

[1] All tumor masks and ground truth labels used for quantitative evaluation were created and verified by the experienced radiologists, as in the original public source datasets [39] (see Table 2 for more details).

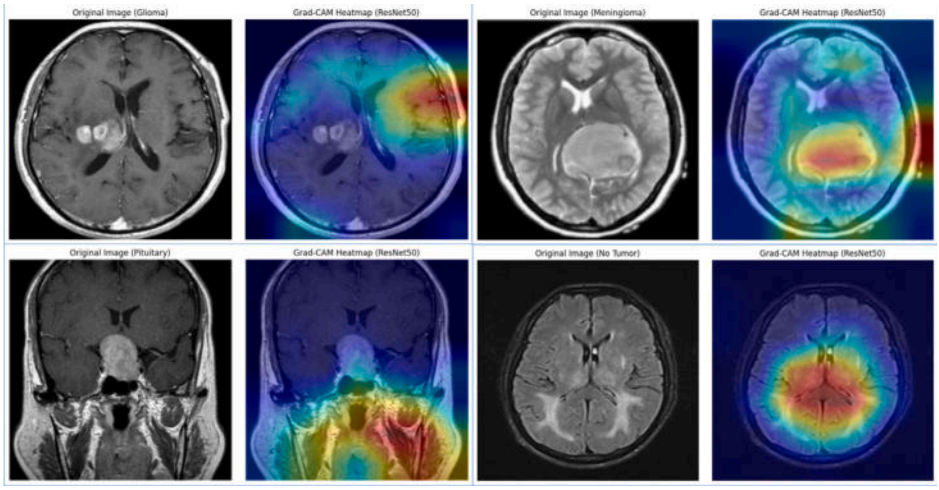


Fig. 11. Grad-CAM heatmaps generated using ResNet50 for four tumor classes: Glioma, Meningioma, Pituitary, and No Tumor. Each pair shows the original MRI image (left) and the corresponding Grad-CAM overlay (right), highlighting class-discriminative regions that influenced the model's prediction.

highlighted attention regions and the true anatomical locations of tumors, providing objective evidence that the Grad-CAM visual explanations are clinically meaningful and trustworthy. This rigorous overlap analysis directly quantifies how well the XAI-based visualizations align with radiological expertise, complementing prediction confidence with spatial interpretability.

8. Key observations and limitations

This study highlights how model architecture, dataset characteristics, augmentation strategies, and interpretability interact in brain tumor classification. SVM emerged as the best classical ML model, reaching up to 97.2% of DL accuracy when trained with AGAN-augmented data. Synthetic samples notably improved minority class representation, boosting class-specific F1-Scores and simultaneously enhancing overall accuracy without introducing new imbalances.

DL models such as ResNet50 and VGG16 achieved the highest overall performance. However, their advantage decreased significantly compared to ML models augmented with AGAN. On smaller or visually ambiguous datasets, the strengths of DL diminished due to overfitting, bringing models such as SVM and XGB to near-parity with DL models in performance. AGAN augmentation effectively addressed class imbalance (as evidenced by balanced distributions and increased F1-scores). The benefit varied by class: minority classes such as MTumour and PTumour saw the greatest improvement, but complex classes like GTumour remained challenging for all model types.

Classical ML models, particularly SVM, KNN, and LR, offered up to 7000× faster training and required over 240× less memory than DL models, making them highly suitable for resource-constrained clinical deployment. Visualizations (t-SNE, UMAP) showed that GAN-generated samples clustered closely with the real data, confirming their semantic validity. XAI tools (SHAP for SVM; LIME and Grad-CAM for DL) improved interpretability. Grad-CAM heatmaps were sometimes diffuse for uncertain or synthetic samples, and the clinical relevance of all interpretability outputs remains unverified by expert review.

Despite robust results, limitations persist: classification performance still varied across datasets due to differences in resolution, noise, and label consistency. The AGAN's ability to reproduce complex tumor morphology, especially for GTumour, was inconsistent. While the integrated XAI framework advances explainability, its clinical utility remains to be validated. Achieving clinically viable AI will require technical improvements, consistent, class-sensitive performance, and close collaboration with medical experts to ensure interpretability and reliability in deployment.

9. Conclusion

This study pursued four key objectives: (i) improving minority-class tumor detection in imbalanced MRI datasets, (ii) integrating interpretability into data augmentation, (iii) benchmarking classical ML and DL models under robust conditions, and (iv) ensuring computational efficiency for clinical deployment.

Our proposed XAI-guided AGAN framework successfully achieved these goals. Minority-class F1-scores improved by 19.7% over the baseline average, demonstrating the efficacy of targeted sample synthesis for class balancing. By incorporating SHAP, LIME, and Grad-CAM scores into GAN training, we established a transparent, clinically guided augmentation process that overcomes conventional "black-box" limitations. Benchmarking across four MRI datasets showed that ML models enhanced with XAI-driven augmentation can achieve 97.2% of state-of-the-art DL accuracy while offering up to 7000-fold computational efficiency gains and over 240× lower memory requirements than DL models. Both quantitative and qualitative interpretability analyses confirmed that the generated samples align with medically meaningful reasoning patterns. In summary, this work highlights the value of efficient, explainable ML pipelines in settings where cutting-edge DL is not feasible. We believe that our XAI-guided augmentation strategy will serve as a foundation for ongoing research into deployable, trustworthy screening tools in healthcare.

Future research should expand along the three promising directions. First, the XAI-GAN feedback mechanism can be extended to multimodal medical imaging (combining MRI, CT, and PET data) to leverage complementary diagnostic information while maintaining interpretability. Second, integrating our AGAN framework within federated learning environments to enable privacy-preserving, multi-institutional synthetic data generation that preserves local data sovereignty while improving global model robustness. Third, developing uncertainty-aware synthetic sample filtering that combines multiple XAI metrics with confidence estimation to create more robust quality control mechanisms for clinical deployment.

CRediT authorship contribution statement

Anbazhagan Mahadevan: Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Radha Reddy:** Writing – review & editing, Supervision, Formal analysis. **Ujwal Nadimpalli:** Writing – original draft, Investigation, Formal analysis. **Mahammad Khaji:** Writing – original draft, Investigation. **Bhavana Nannuru:** Writing – original draft, Investigation. **Aateesh Akkina:** Writing – original draft, Methodology, Investigation.

Code availability

The source code is not publicly released at this stage. However, it is available from the corresponding author upon reasonable request for academic research purposes.

Ethics statement

All data used in this study were obtained from publicly available and fully de-identified datasets. Each dataset was released under open data sharing arrangements for research purposes. There was no direct involvement of human subjects, and the authors did not access or process private or sensitive information.

Tumor segmentation masks and ground truth labels used for quantitative evaluation (e.g., Intersection over Union (IoU) calculations for Grad-CAM validation) were provided by the dataset creators or obtained from prior expert annotations as documented in the respective data sources. The authors of the present study did not conduct any new annotations or assessments.

This study did not involve new human subjects and used only a secondary, fully de-identified, publicly available dataset. As a result, it was exempt from the institutional ethical review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Notations

See Table A.8

Data availability

All datasets used in this study are publicly available and fully de-identified:

• **Dataset 1:** Kaggle — Brain Tumor Classification (MRI). Available at: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri

• **Dataset 2:** Figshare — Brain Tumor Dataset. Available at: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

• **Dataset 3:** Kaggle — Crystal Clean: Brain Tumors MRI Dataset. Available at: https://www.kaggle.com/datasets/mohammadhossein77/brain-tumors-dataset

• **Dataset 4:** Kaggle — Brain Tumour Classification. Available at: https://www.kaggle.com/datasets/rishiksaisanthosh/brain-tumour-classification

All access links and detailed dataset specifications are listed in Table 2 in the main manuscript. There are no access restrictions for these publicly available datasets. Derived data, such as dimensionally reduced features (PCA components), GAN-generated synthetic samples, and interpretability outputs (SHAP values and Grad-CAM visualizations), will be made available upon reasonable request.

Table A.8
Notation table: Key symbols and terms.

Notation/Symbol	Meaning
D	Original training MRI dataset
N	Number of images in the dataset
$D_{original}$	Set of original training data
D_{aug}	Augmented dataset with real + synthetic samples
x_i	ith MRI image/sample
y_i	Label of the ith sample
c	Class index (tumor class)
n_c	Number of samples in class c
C, $C_{minority}$	Set of all classes, set of minority classes
z	Latent noise vector for GAN generator
G_c	GAN generator for class c
D_c	GAN discriminator for class c
$G_c(z)$	Generated (synthetic) sample for class c from generator
$Q_{XAI}(x_i)$	XAI quality/interpretability score for sample x_i
$Q_{GC}(x_j)$	Grad-CAM based interpretability score for CNN models
$Q_{SHAP}(x_j)$	SHAP-based interpretability score (for ML models)
α	XAI Weighting
$\mathcal{X}^t_{cand}, \mathcal{X}^t_{filtered}$	Set of candidate/generated samples and filtered high-quality samples at iteration t
$\mathcal{X}_{GAN,c}$	Final set of retained GAN-generated samples for class c
$\mathcal{X}_{augmented}$	Final overall augmented training set
τ_{XAI}	XAI interpretability quality threshold
$V(D,G)$	GAN objective function (value function)
λ_{XAI}	Weight (penalty) for interpretability in the loss
γ_c	Class-wise imbalance penalty
ξ	Retention rate (fraction of synthetic samples retained above threshold)
$\mathbf{I}$	Image tensor (MRI image)
k	Number of PCA principal components retained
$\sigma_{R/G/B}, \mu_{R/G/B}$	Channel-wise std. dev. and mean for image normalization
f	Nonlinear activation function (GAN, SHAP, etc.)
$\mathcal{L}$	Generic loss function (e.g., for LIME, SHAP, GAN)
C_{base}	Baseline classifier (e.g., SVM, CNN, etc.)
T	Number of AGAN feedback iterations
T_{train}	Training time
Mem	Peak memory usage during training

References

[1] Park J, Park YG. Brain tumor rehabilitation: symptoms, complications, and treatment strategy. Brain Neurorehabil 2022;15(3):e25.
[2] Santhosh TRS, Mohanty SN, Pradhan NR, Khan T, Derbali M. Neurovision: A deep learning driven web application for brain tumour detection using weight-aware decision approach. Digit Health 2025;11:20552076251333195.
[3] Poernama AI, Soesanti I, Wahyunggoro O. Feature extraction and feature selection methods in classification of brain MRI images: a review. In: 2019 international biomedical instrumentation and technology conference, vol. 1, IEEE; 2019, p. 58–63.
[4] Dogan Y. A novel pruning-enhanced hybrid approach for efficient and accurate brain tumor diagnosis. Biomed Signal Process Control 2026;112:108466.
[5] Öter E, Doğan Y. A comparative study on data balancing methods for Alzheimer's disease classification. Çukurova Üniversitesi Mühendislik Fakültesi Derg 2024;39(2):489–501.
[6] Missaoui R, Hechkel W, Saadaoui W, Helali A, Leo M. Advanced deep learning and machine learning techniques for MRI brain tumor analysis: A review. Sensors 2025;25(9):2746.
[7] Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decis Anal J 2022;3:100071.
[8] Asiri AA, Khan B, Muhammad F, Alshamrani HA, Alshamrani KA, Irfan M, Alqhtani FF, et al. Machine learning-based models for magnetic resonance imaging (MRI)-based brain tumor classification. Intell Autom Soft Comput 2023;36:299–312.
[9] Kumar AA, Kesavadas C. Potential of MRI in clinical medicine. In: Multimodal biomedical imaging techniques. Springer; 2025, p. 271–301.
[10] Martucci M, Russo R, Schimperna F, D'Apolito G, Panfili M, Grimaldi A, Perna A, Ferranti A, Varcasia G, Giordano C, et al. Magnetic resonance imaging of primary adult brain tumors: State of the art and future perspectives. Biomedicines 2023;11:364.
[11] Deepak S, Ameer P. Brain tumor categorization from imbalanced MRI dataset using weighted loss and deep feature fusion. Neurocomputing 2023;520:94–102.
[12] Berisha V, Krantsevich C, Hahn P, Hahn S, Dasarathy G, Turaga P, Liss J. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021;4(1):153.

[13] Nhlapho W, Atemkeng M, Brima Y, Ndogmo J-C. Bridging the gap: exploring interpretability in deep learning models for brain tumor detection and diagnosis from MRI images. Information 2024;15(4):182.
[14] Solanki S, Singh UP, Chouhan SS. Brain tumor classification using ML and DL approaches. In: 2023 IEEE 5th international conference on cybernetics, cognition and machine learning applications. IEEE; 2023, p. 204–8.
[15] Kuppusamy P, Kodavaluru VS, Bogar SM. Brain tumor classification using optimal features and ensemble learning algorithms. In: 2023 first international conference on advances in electrical, electronics and computational intelligence. IEEE; 2023, p. 1–8.
[16] Jayanthi R, Christinal AH, Hephzibah R, Shekinah T, Bajaj C, Chandy DA. A novel perspective on brain tumor classification using hybrid algorithm. In: 2023 international conference on computer, electronics & electrical engineering & their applications. IEEE; 2023, p. 1–4.
[17] Anilkumar B, Kumar NP, Sowmya K. Mr brain tumour classification using a deep ensemble learning technique. In: 2023 5th biennial international conference on nascent technologies in engineering. IEEE; 2023, p. 1–6.
[18] Anitha G, Usman A, Ahmed S, Nasser H. Detection of brain tumor using machine learning classifiers–a comparative study. In: 2022 OPJU international technology conference on emerging technologies for sustainable development. IEEE; 2023, p. 1–6.
[19] Mishra SK, Kumar D, Kumar G, Kumar S. Multi-classification of brain MRI using EfficientNet. In: 2022 international conference for advancement in technology. IEEE; 2022, p. 1–6.
[20] Islam S, Udhayakumar R, Kar S, Aswal US, Saini DKJB, et al. Enhancing brain tumor detection classification by computational intelligence. In: 2023 second international conference on electronics and renewable systems. IEEE; 2023, p. 1044–9.
[21] Kibriya H, Masood M, Nawaz M, Rafique R, Rehman S. Multiclass brain tumor classification using convolutional neural network and support vector machine. In: 2021 Mohammad Ali Jinnah university international conference on computing. IEEE; 2021, p. 1–4.
[22] Ferdous GJ, Sathi KA, Hossain MA. Application of hybrid classifier for multi-class classification of MRI brain tumor images. In: 2021 5th international conference on electrical engineering and information communication technology. IEEE; 2021, p. 1–6.
[23] UmaMaheswaran S, Deivasigamani S, Joshi K, Verma D, Rajamani SK, Ross DS. Computational intelligence approach to improve the classification accuracy of brain tumor detection. In: 2022 11th international conference on system modeling & advancement in research trends. IEEE; 2022, p. 1192–6.
[24] Upadhyay U, Sudarshan VP, Awate SP. Uncertainty-aware gan with adaptive loss for robust mri image enhancement. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 3255–64.
[25] Han C, Rundo L, Araki R, Nagano Y, Furukawa Y, Mauri G, Nakayama H, Hayashi H. Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection. IEEE Access 2019;7:156966–77.
[26] Neelima G, Chigurukota DR, Maram B, Girirajan B. Optimal DeepMRSeg based tumor segmentation with GAN for brain tumor classification. Biomed Signal Process Control 2022;74:103537.
[27] Nag A, Mondal H, Hassan MM, Al-Shehari T, Kadrie M, Al-Razgan M, Alfakih T, Biswas S, Bairagi AK. TumorGANet: A transfer learning and generative adversarial network-based data augmentation model for brain tumor classification. IEEE Access 2024.
[28] Karpakam S, Kumareshan N. Enhanced brain tumor detection and classification using a deep image recognition generative adversarial network (DIR-gan): A comparative study on MRI, X-ray, and FigShare datasets. Neural Comput Appl 2025;1–28.
[29] Nazir MI, Akter A, Wadud MAH, Uddin MA. Utilizing customized CNN for brain tumor prediction with explainable AI. Heliyon 2024;10(20).
[30] Gaur L, Bhandari M, Razdan T, Mallik S, Zhao Z. Explanation-driven deep learning model for prediction of brain tumour status using MRI image data. Front Genet 2022;13:822666.
[31] Lin LP, Seow ZH. Classifying brain tumours: a deep learning approach with explainable AI. In: Proceedings of the 2024 14th international conference on biomedical engineering and technology. 2024, p. 101–7.
[32] Ariful Islam M, Mridha M, Safran M, Alfarhood S, Mohsin Kabir M. Revolutionizing brain tumor detection using explainable AI in MRI images. NMR Biomed 2025;38(3):e70001.
[33] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP. Preparing medical imaging data for machine learning. Radiology 2020;295(1):4–15.
[34] Bhuvaji S, Kadam A, Bhumkar P, Dedge S, Kanchan S. Brain tumor classification (MRI). Kaggle; 2020, http://dx.doi.org/10.34740/KAGGLE/DSV/1183165, URL https://www.kaggle.com/dsv/1183165.
[35] Cheng J. Brain tumor dataset. figshare; 2015, (No Title).
[36] Hashemi SMH. Crystal clean: Brain tumors MRI dataset. Kaggle; 2023, [Accessed 09 May 2023].
[37] Poernama AI, Soesanti I, Wahyunggoro O. Feature extraction and feature selection methods in classification of brain MRI images: a review. In: 2019 international biomedical instrumentation and technology conference, vol. 1, IEEE; 2019, p. 58–63.
[38] Liang DD, Liang DD, Pomeroy MJ, Gao Y, Kuo LR, Li LC. Examining feature extraction and classification modules in machine learning for diagnosis of low-dose computed tomographic screening-detected in vivo lesions. J Med Imaging 2024;11(4). 044501–044501.
[39] Hammad M, ElAffendi M, Ateya AA, Abd El-Latif AA. Efficient brain tumor detection with lightweight end-to-end deep learning model. Cancers 2023;15(10):2837.