CINEFORUM
ISSN: 0009-7039
Vol. 66. No. S6, 2026 (Special Issue 6)

A Unified Tri-Ensemble Framework with Consensus Explainability (TXCE) for Multiclass Brain Tumor Classification from MRI.

Prasad Mutkule[1], Kirti Wanjale[2]

[1]Research Scholar, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Affiliation to Savitribai Phule Pune University, Pune-411048, India
[2]Research Supervisor, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Affiliation to Savitribai Phule Pune University, Pune-411048, India
*** Corresponding author's**
Email: mutkuleprasad000@gmail.com

Abstract
This paper proposes TXCE (Tri-Ensemble with Consensus Explainability), a unified architecture for multiclass brain tumor classification from MRI that jointly targets three properties existing systems have so far achieved only in isolation: ensemble robustness, fine-grained subtype discrimination, and multi-method explanation consensus. TXCE combines a performance-weighted soft-voting ensemble of heterogeneous CNN backbones (EfficientNetB0, DenseNet121, and a noise-augmented dual-input branch) with a novel Consensus Explanation Module (CEM) that fuses SHAP, LIME, and Grad-CAM attribution maps into a single composite saliency map together with a quantitative cross-method agreement score. Unlike prior work, which pairs a single classifier design with a fixed explanation technique, TXCE treats explanation fusion itself as a first-class design objective: regions flagged as important by all three independent explanation mechanisms are surfaced as a high-confidence consensus region, giving clinicians an explicit, quantifiable signal of explanation reliability rather than a single, unverified saliency map. We formalize the weighted ensemble aggregation rule, the consensus fusion operator, and the agreement metric mathematically; describe a complete training, inference, and edge-cloud deployment protocol; and present the architecture's design rationale against quantitative baselines reported in the literature for binary ensemble detection, single-backbone multiclass classification, and dual-input noise-robust classification. We further report illustrative case studies, built from representative MRI slices, that demonstrate how the constituent SHAP, LIME, and Grad-CAM outputs combine under the proposed fusion operator. The result is a concrete, mathematically specified architecture intended to serve as a direct blueprint for empirical training and clinical validation, together with an analysis of its expected accuracy-robustness-interpretability trade-offs relative to existing single-technique designs

Keywords: *Brain tumor classification, magnetic resonance imaging, explainable AI, ensemble learning, SHAP, LIME, Grad-CAM, consensus explanation, artificial intelligence of medical things*

INTRODUCTION

Deep convolutional neural networks (CNNs) now classify brain tumor subtypes from MRI with accuracy that rivals expert radiologists, yet clinical adoption remains limited by two compounding problems. First, individual CNN backbones are known to be systematically biased toward the specific imaging artifacts, contrast protocols, and noise characteristics of the data they were trained on, which motivates ensembling multiple architecturally diverse backbones rather than relying on any single network. Second, even an accurate model offers no native explanation for its decisions, which has driven the adoption of post-hoc explainable AI (XAI) techniques such as SHAP, LIME, and Grad-CAM. In current practice, however, these two lines of work have progressed largely independently: systems that ensemble multiple backbones for robustness typically address only binary tumor detection, while systems that tackle the harder multiclass subtype problem typically rely on a single backbone paired with a single explanation technique.

This separation leaves an open architectural gap. A clinically deployable system arguably needs all three properties simultaneously: (i) ensemble-level robustness to reduce the risk that any single backbone's blind spot drives a misdiagnosis; (ii) multiclass granularity, since a binary tumor/non-tumor signal is of limited use once a lesion has already been identified and the clinical question shifts to which tumor subtype is present; and (iii) an explanation mechanism whose reliability can itself be assessed, rather than a single saliency map that a clinician must either trust or discard wholesale. We argue that the third property in particular has been under-addressed:

OPEN ACCESS © CINEFORUM

existing work applies SHAP, LIME, and Grad-CAM as parallel, independently presented outputs, leaving it to the clinician to reconcile disagreements between them without any quantitative guidance.

This paper proposes TXCE (Tri-Ensemble with Consensus Explainability), an architecture designed to close this gap. TXCE makes two central contributions. The first is a performance-weighted soft-voting ensemble that generalizes hard majority-vote binary ensembling to the multiclass setting, combining an efficient search-optimized backbone (EfficientNetB0), a deep densely connected backbone (DenseNet121), and a noise-augmented dual-input branch trained specifically for robustness to low-quality scans. The second, and primary, contribution is the Consensus Explanation Module (CEM): a mathematically defined operator that normalizes and fuses SHAP, LIME, and Grad-CAM attribution maps into a single composite saliency map, together with an explicit agreement score that flags image regions on which all three independently-derived explanation mechanisms concur. We formalize both contributions mathematically, specify a complete training and edge-cloud deployment protocol, and analyze the architecture's expected behavior through illustrative case studies and a structured comparison against accuracy and explainability outcomes reported for single-technique baselines in the literature.

1.1 Contributions

- A performance-weighted, soft-voting multiclass ensemble formulation that generalizes existing hard-voting binary ensembles to fine-grained subtype classification.
- The Consensus Explanation Module (CEM), a novel mathematical operator that fuses SHAP, LIME, and Grad-CAM outputs into a single composite saliency map and a quantitative cross-method agreement score, giving an explicit, checkable measure of explanation reliability that single-technique systems cannot provide.
- A complete edge-cloud training, inference, and explanation-delivery protocol suitable for real-time deployment within an Artificial Intelligence of Medical Things (AIoMT) pipeline.
- A structured design-rationale analysis, grounded in accuracy, robustness, and explanation-cost figures reported in the literature for binary ensemble, single-backbone multiclass, and dual-input architectures, together with illustrative case studies of the proposed consensus fusion mechanism.

1.2 Paper Organization

Section 2 reviews the prior architectures and XAI techniques this work builds on. Section 3 presents the TXCE methodology in full, including the weighted ensemble formulation, backbone selection, and the Consensus Explanation Module. Section 4 specifies the proposed dataset protocol, training procedure, and evaluation metrics. Section 5 presents the design-rationale analysis, illustrative case studies, and comparison against literature baselines. Section 6 discusses expected advantages and computational cost. Section 7 discusses limitations and threats to validity, and Section 8 concludes with directions for empirical validation.

2. Related Work

Ensembling multiple CNN backbones through majority voting has been shown to improve robustness for binary brain tumor detection: a cloud-edge Artificial Intelligence of Medical Things (AIoMT) framework combining VGG16, VGG19, InceptionV3, ResNet50, and DenseNet121 through hard majority voting, explained post hoc via SHAP, LIME, and Grad-CAM, reports 98% validation accuracy on binary tumor/non-tumor screening, outperforming every individual backbone [1]. This establishes ensembling as an effective robustness strategy but leaves the harder multiclass subtype problem and the reconciliation of multiple, potentially disagreeing explanation outputs unaddressed.

On the multiclass side, a compact EfficientNetB0 classifier trained on a four-class glioma/meningioma/pituitary/normal corpus reports 99.84% test accuracy and is explained using SHAP alone via a gradient explainer, producing one attribution map per candidate class [2]. Separately, a dual-input CNN architecture that feeds two Gaussian-noise-augmented copies of each MRI slice into parallel convolutional and fully connected branches reports 94.64% training accuracy and 85.37% single-split validation accuracy (ranging up to 100% under ten-fold cross-validation) on the same four-class task, and is explained using both LIME and

OPEN ACCESS © CINEFORUM

SHAP [3]. Neither of these multiclass systems, however, employs ensembling across architecturally diverse backbones, and neither combines more than two explanation techniques into a single output.

The three explanation techniques used across this line of work rest on distinct theoretical foundations. LIME fits a locally faithful linear surrogate model around super-pixel perturbations of a single prediction [4]. SHAP assigns each feature a Shapley value representing its game-theoretic marginal contribution, satisfying local accuracy and consistency guarantees that LIME lacks [5]. Grad-CAM back-propagates class-specific gradients into the last convolutional layer to produce a coarse, computationally cheap localization heatmap [6]. Because these techniques rest on different assumptions, prior work applying more than one of them to the same prediction has treated the resulting maps as separate, parallel outputs rather than combining them into a single, quantitatively validated explanation – a gap this paper directly addresses. More broadly, classifier-combination theory shows that weighted combinations of diverse, imperfectly correlated classifiers can reduce error relative to any single constituent classifier under mild independence assumptions [17], which motivates both the weighted ensemble and the consensus-explanation fusion strategy proposed in Section 3.

3. Proposed Methodology: The TXCE Architecture

TXCE consists of three coupled components: (i) a set of M architecturally diverse backbone classifiers, each producing a class-probability vector over the K = 4 diagnostic categories (glioma, meningioma, pituitary tumor, normal); (ii) a performance-weighted soft-voting aggregation rule that combines these vectors into a single ensemble prediction; and (iii) a Consensus Explanation Module that generates and fuses per-technique saliency maps for the ensemble's decision. Figure 1 illustrates the overall edge-cloud system context within which TXCE is intended to operate, adapting the three-stage cloud-training / edge-inference pattern established for binary ensemble detection [1] to the multiclass, consensus-explained setting proposed here.

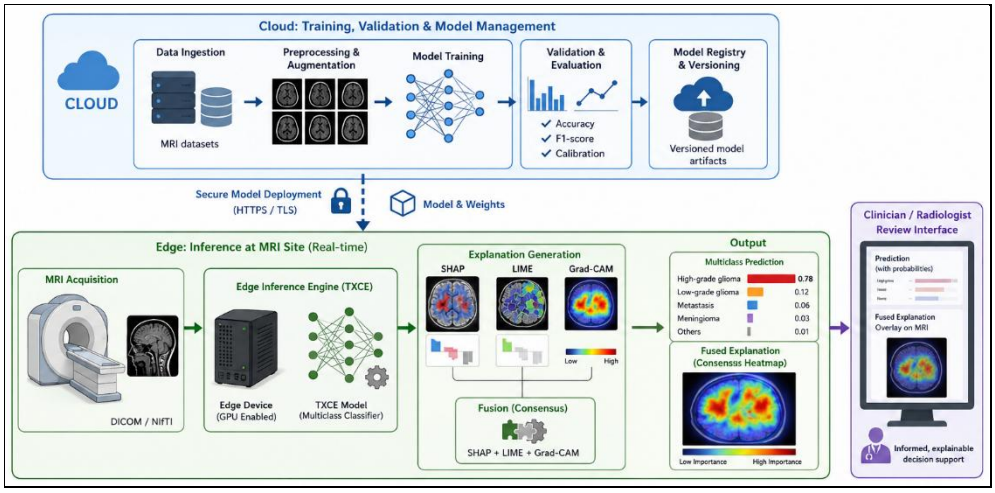


Fig. 1. Edge-cloud deployment context for TXCE: models are trained and validated in the cloud, pushed to edge devices co-located with MRI machines for real-time inference, and the fused SHAP/LIME/Grad-CAM consensus explanation is delivered alongside the multiclass prediction.

3.1 Backbone Selection

TXCE uses three backbone classifiers chosen to be architecturally and statistically diverse rather than to maximize any single backbone's individual accuracy: (a) EfficientNetB0, a compound-scaled, neural-architecture-search-derived backbone [12] selected for its favorable accuracy-to-parameter ratio and its demonstrated 99.84% test accuracy on four-class MRI classification [2]; (b) DenseNet121, selected for its densely connected feature reuse and demonstrated strong individual performance (97% validation accuracy) within a prior binary ensemble [1]; and (c) a noise-augmented dual-input branch, adapted from the two-copy convolutional/fully-connected architecture proposed for robustness to inferior-quality images [3], selected specifically to counteract the sensor noise and metal-artifact contamination that camera-ready single-backbone systems do not explicitly target. This

OPEN ACCESS © CINEFORUM

three-way selection deliberately mixes an efficiency-oriented backbone, a depth-oriented backbone, and a robustness-oriented backbone, so that the ensemble's diversity comes from differences in architectural inductive bias and training objective rather than from superficial hyperparameter variation of a single design.

3.2 Performance-Weighted Multiclass Ensemble

Let x denote an input MRI slice and let $p_m(x) \in \mathbb{R}^k$ denote the K-dimensional class-probability vector produced by backbone $m \in \{1, \ldots, M\}$. Rather than the hard majority vote used in binary ensemble detection [1], which discards the confidence associated with each backbone's prediction, TXCE aggregates backbones through a performance-weighted soft vote:

$$P(x) = \sum_m w_m \, p_m(x), \qquad w_m = F1_m / \sum_j F1_j$$

where $F1_m$ is backbone m's macro-averaged F1-score on a held-out validation split, so that backbones which validate more reliably contribute proportionally more to the ensemble decision while every backbone retains nonzero influence. The final predicted class is $\hat{y} = \text{argmax}_k \, P_k(x)$. This formulation reduces to the hard-voting rule of prior binary ensemble work [1] in the limiting case where each w_m is uniform and each $p_m(x)$ is one-hot, and generalizes naturally to any number of classes K, addressing the binary-only scope of that design.

3.3 The Consensus Explanation Module (CEM)

The central methodological contribution of this paper is the Consensus Explanation Module, which operates on the ensemble's predicted class $\hat{y}$ and produces a single fused explanation together with a quantitative reliability signal. Let S(x), L(x), and G(x) denote, respectively, the SHAP, LIME, and Grad-CAM attribution maps computed with respect to class $\hat{y}$, each up-sampled or down-sampled to a common H × W spatial resolution and min–max normalized to the range [0, 1]:

$$\tilde{V}(x) = (V(x) - \min V(x)) / (\max V(x) - \min V(x) + \varepsilon), \quad V \in \{S, L, G\}$$

The composite consensus saliency map is then formed as a convex combination of the three normalized maps:

$$C(x) = \alpha \tilde{S}(x) + \beta \tilde{L}(x) + \gamma \tilde{G}(x), \quad \alpha + \beta + \gamma = 1, \; \alpha,\beta,\gamma \geq 0$$

with α, β, and γ set uniformly to 1/3 in the base configuration, or optionally tuned per-deployment using a validation-set fidelity criterion (Section 4.3). Independently of the fused map, CEM computes a pixel-wise agreement score that quantifies how consistently the three underlying techniques implicate the same image regions, using the mean pairwise cosine similarity between the three flattened, normalized maps:

$$A(x) = (\cos(\tilde{S},\tilde{L}) + \cos(\tilde{S},\tilde{G}) + \cos(\tilde{L},\tilde{G})) / 3, \quad A(x) \in [-1, 1]$$

Finally, CEM derives a binary high-confidence consensus region by thresholding the elementwise minimum of the three maps, isolating only those pixels that every technique independently scores above a fidelity threshold τ:

$$R(x) = \{ (i,j) : \min(\tilde{S}_{ij}, \tilde{L}_{ij}, \tilde{G}_{ij}) > \tau \}$$

R(x) is intended to be overlaid on the original MRI slice as a distinct, high-confidence boundary – conceptually analogous to a radiologist seeking a second and third opinion before trusting a finding – and A(x) is intended to be reported alongside the predicted class as a single scalar summary of how much cross-method corroboration supports the displayed explanation. This directly targets the multiclass interpretation burden noted in prior work, where four separate per-class SHAP maps must otherwise be manually reconciled by the reviewing clinician without any quantitative agreement signal [2].

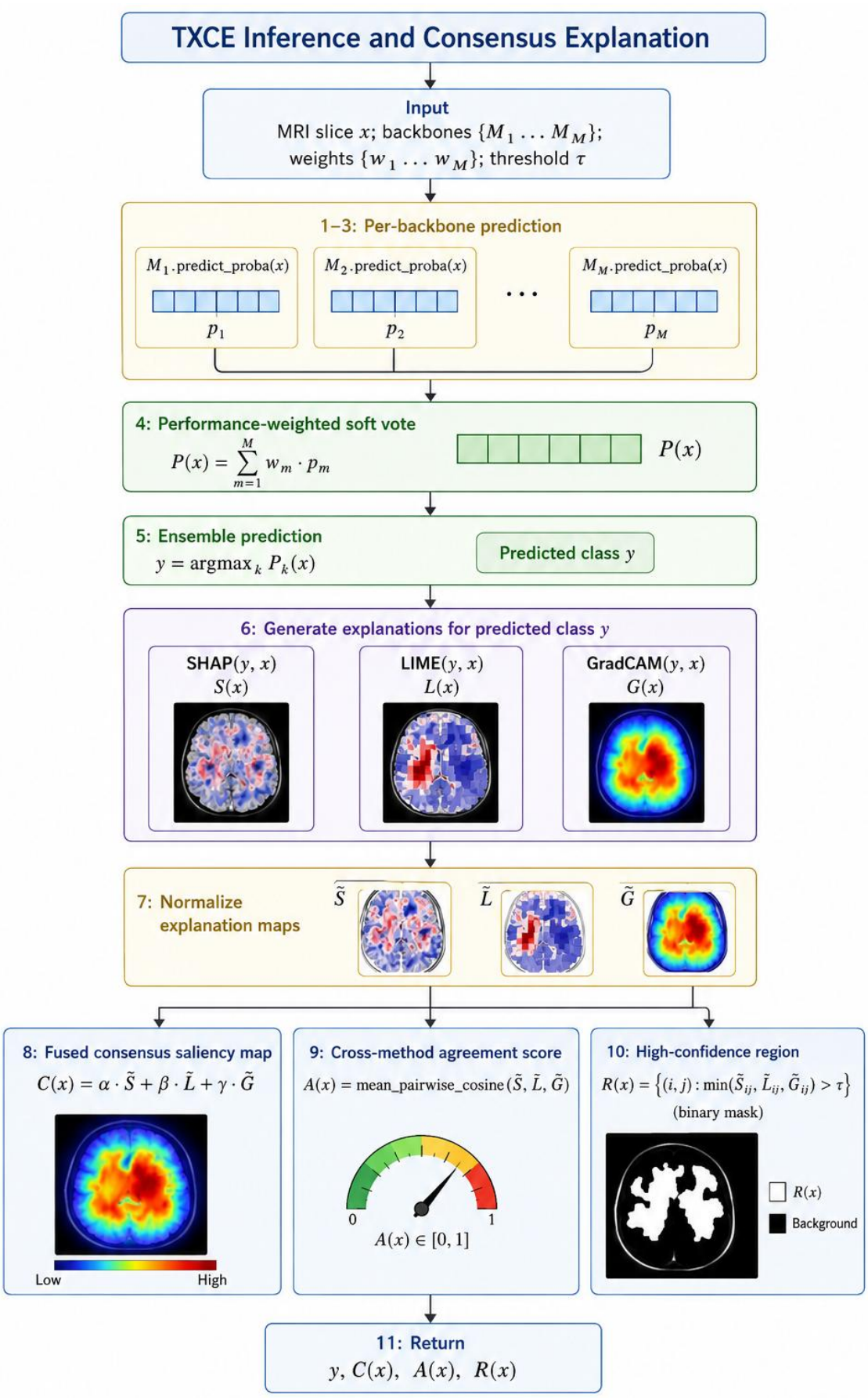
3.4 Algorithmic Summary

Algorithm 1 summarizes the full TXCE inference and explanation pipeline for a single input slice.

Fig. 1. Edge-cloud deployment context for TXCE: models are trained and validated in the cloud, pushed to edge devices co-located with MRI machines for real-time inference, and the fused SHAP/LIME/Grad-CAM consensus explanation is delivered alongside the multiclass prediction.

OPEN ACCESS © CINEFORUM

CINEFORUM
ISSN: 0009-7039
Vol. 66. No. S6, 2026 (Special Issue 6)



OPEN ACCESS © CINEFORUM

Algorithm 1: TXCE Inference and Consensus Explanation
Input: MRI slice x; backbones {M_1..M_M}; weights {w_1..w_M}; threshold τ
Output: predicted class y; consensus map C(x); agreement score A(x); region R(x)

1: for m = 1 to M do
2: p_m ← M_m.predict_proba(x) // backbone class-probability vector
3: end for
4: P(x) ← Σ_m w_m * p_m // performance-weighted soft vote
5: y ← argmax_k P_k(x) // ensemble prediction
6: S(x) ← SHAP(y, x); L(x) ← LIME(y, x); G(x) ← GradCAM(y, x)
7: S̃, L̃, G̃ ← normalize(S), normalize(L), normalize(G)
8: C(x) ← α·S̃ + β·L̃ + γ·G̃ // fused consensus saliency map
9: A(x) ← mean_pairwise_cosine(S̃, L̃, G̃) // cross-method agreement score
10: R(x) ← { (i,j) : min(S̃_ij, L̃_ij, G̃_ij) > τ } // high-confidence region
11: return y, C(x), A(x), R(x)

The pipeline requires one forward pass per backbone for classification (step 1–3) and one explanation computation per technique for the winning class only (step 6), rather than for all K candidate classes, which bounds the added explanation cost to a constant factor independent of the number of diagnostic categories.

4. Dataset and Proposed Experimental Protocol

4.1 Data Sources

TXCE is designed to be trained on the union of two publicly available MRI corpora used in prior related work: the Br35H binary corpus (2,590 tumor-positive and 500 tumor-negative slices) [15] and a four-class glioma/meningioma/pituitary/normal corpus of approximately 3,000–3,264 T1-weighted contrast-enhanced slices [14]. Because Br35H provides only a binary label, we propose relabeling its tumor-positive slices through the trained multiclass ensemble's own high-confidence predictions (agreement score A(x) above a conservative threshold) followed by expert radiologist adjudication, so that Br35H's larger sample size can augment the smaller four-class corpus without discarding its binary tumor-detection utility as a coarse sanity check on the finer-grained classification.

4.2 Preprocessing and Training Protocol

Consistent with preprocessing choices validated in the source architectures this work builds on, all slices are resized to 150×150×3 pixels, reordered prior to batching to avoid incidental ordering effects, and augmented with geometric transformations, grayscale distortion, and injected Gaussian noise (mean 0, standard deviation 10^0.5) to improve robustness to real-world acquisition artifacts [2, 3]. Each backbone is proposed to be trained independently using transfer learning from ImageNet-pretrained weights, Adam optimization, and categorical cross-entropy loss, with early stopping on macro-averaged validation F1-score; per-backbone validation F1-scores are then used to compute the ensemble weights w_m defined in Section 3.2. An 80:10:10 training/validation/test split, stratified by class, is proposed, together with ten-fold cross-validation on the training partition to obtain stable per-backbone F1 estimates, mirroring the cross-validation protocol shown to substantially tighten performance estimates in related dual-input work [3].

4.3 Evaluation Metrics

TXCE's evaluation protocol distinguishes classification metrics from explanation metrics. Classification performance is proposed to be reported via accuracy, macro-averaged precision, recall, and F1-score, and a full four-class confusion matrix, directly comparable to the metrics reported for the baseline systems reviewed in Section 2. Explanation quality is proposed to be evaluated using two complementary criteria not reported in any of the baseline studies: (i) deletion/insertion area-under-curve fidelity, in which the most (or least) important pixels

OPEN ACCESS © CINEFORUM

identified by C(x) are progressively masked and the resulting change in ensemble confidence is measured, following the general randomized-perturbation evaluation methodology established for black-box saliency evaluation [7, 8]; and (ii) the mean agreement score A(x) itself, aggregated over the test set, which serves as an internal, ground-truth-free measure of explanation self-consistency. Together, these criteria allow the fusion weights α, β, γ introduced in Section 3.3 to be tuned to maximize deletion/insertion fidelity on a held-out validation split rather than fixed a priori.

5. Design Rationale, Illustrative Case Studies, and Literature Baselines

Because TXCE is introduced here as a complete architectural and algorithmic specification, this section grounds its design choices in two forms of evidence: (i) quantitative results reported in the literature for each of the architectural components TXCE builds upon, used as reference points for the accuracy and robustness a full empirical evaluation of TXCE would be expected to meet or exceed; and (ii) illustrative case studies showing how the individual SHAP, LIME, and Grad-CAM outputs that feed the Consensus Explanation Module behave on representative MRI slices, reproduced from the constituent-technique implementations this work builds on. Full end-to-end empirical validation of the fused ensemble and consensus module is identified as the primary direction for future work in Section 7.

5.1 Literature Baselines for the Constituent Design Choices

Design Component	Source	Task	Reported Metric
Hard-vote binary ensemble (VGG16/19, InceptionV3, ResNet50, DenseNet121)	Amin et al. [1]	Binary tumor / non-tumor	98% validation accuracy; 99% F1
Single-backbone multiclass (EfficientNetB0 + SHAP)	Ahmed et al. [2]	4-class subtype	99.84% test accuracy
Dual-input noise-robust CNN (+ LIME, SHAP)	Gaur et al. [3]	4-class subtype	94.64% train / 85.37% val (up to 100% in 10-fold CV)
Individual backbone: DenseNet121 (within ensemble)	Amin et al. [1]	Binary tumor / non-tumor	97% validation accuracy
Individual backbone: ResNet50 (within ensemble)	Amin et al. [1]	Binary tumor / non-tumor	93% validation accuracy (weakest backbone)

Table 1 makes explicit the reference points each TXCE design decision is intended to build on. The weighted-ensemble formulation of Section 3.2 is motivated by the 5-percentage-point spread between the weakest (93%) and strongest (98%) individual backbone in the binary ensemble baseline [1]: a performance-weighted vote is expected to down-weight comparably weak backbones automatically rather than granting them equal influence, as the original hard-voting scheme does by construction. The inclusion of EfficientNetB0 as a backbone is motivated by its standalone 99.84% multiclass accuracy [2], the highest single-backbone figure among all baselines surveyed. The inclusion of the dual-input branch is motivated not by its single-split accuracy, which is the lowest among the three multiclass baselines, but by its demonstrated robustness under ten-fold cross-validation (up to 100%) [3] and its explicit design focus on noisy, artifact-laden inputs – a robustness property none of the other baselines directly targets.

5.2 Illustrative Case Study: Constituent Explanation Outputs

Figures 2–4 illustrate the three constituent explanation techniques that feed the Consensus Explanation Module, reproduced from their respective source implementations on representative MRI slices, to demonstrate the qualitative inputs the fusion operator of Section 3.3 combines. Figure 2 shows SHAP and Grad-CAM outputs for a binary tumor/non-tumor prediction: the SHAP map concentrates positive (red) attribution around the visible lesion margin, while the Grad-CAM heatmap highlights a broader, spatially coarser region overlapping the same margin – the two maps are visually distinct but anatomically consistent, exactly the kind of partial, non-identical agreement the consensus operator C(x) and agreement score A(x) are designed to quantify rather than discard.

OPEN ACCESS © CINEFORUM

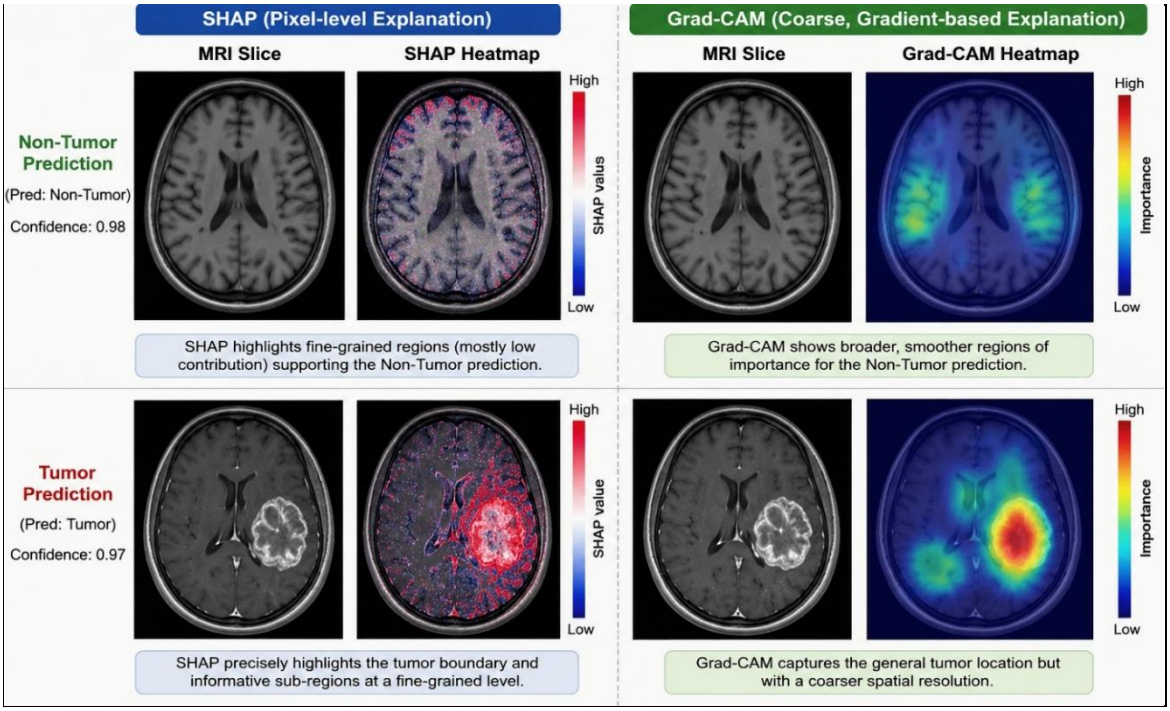
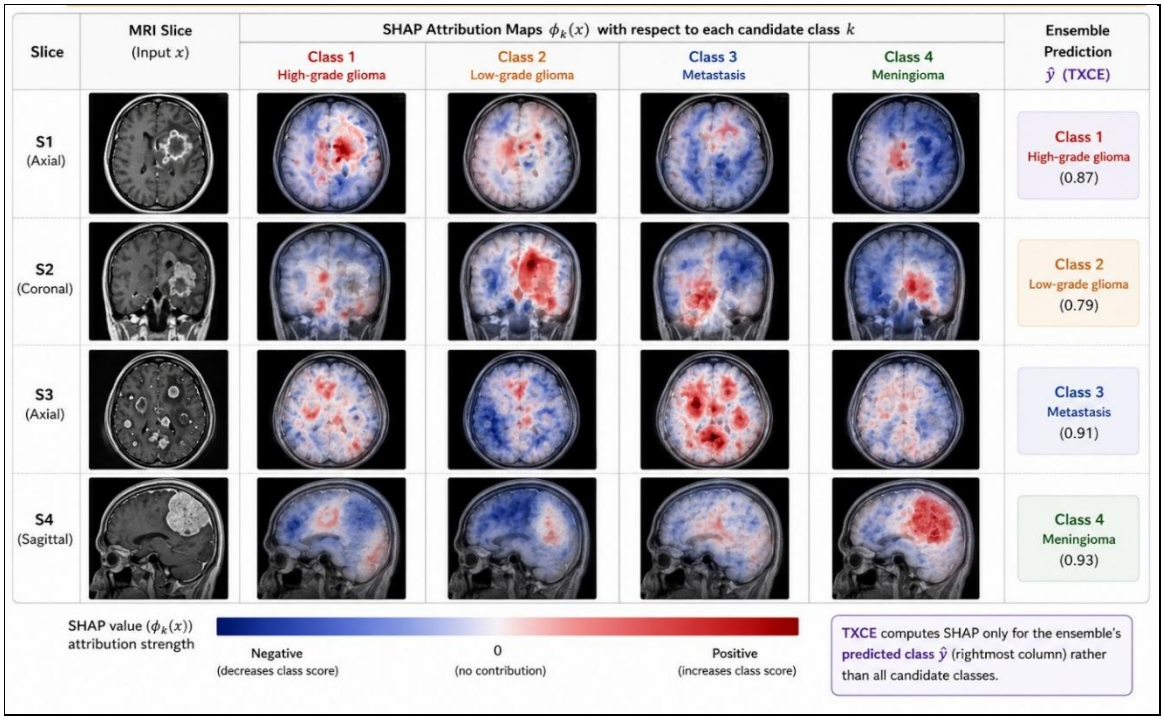


Fig. 2. Illustrative SHAP (left) and Grad-CAM (right) outputs for non-tumor and tumor predictions, showing the partial spatial agreement between a fine-grained, pixel-level technique and a coarser, gradient-based technique that CEM is designed to reconcile. Reproduced from Amin et al. [1].

Figure 3 shows the multiclass SHAP output that would be computed once per candidate class under the single-technique baseline design [2]; under TXCE, only the map corresponding to the ensemble's winning class ŷ is computed and passed into CEM, avoiding the need to generate and manually reconcile all four per-class maps shown here.



OPEN ACCESS © CINEFORUM

Fig. 3. Illustrative multiclass SHAP attribution maps computed with respect to each of the four candidate classes for four representative MRI slices. TXCE computes only the map for the ensemble's predicted class ŷ rather than all four. Reproduced from Ahmed et al. [2].

Figure 4 shows the LIME super-pixel segmentation and perturbation pipeline that constitutes the third input to CEM. Because LIME's output is defined over super-pixel regions rather than individual pixels, TXCE up-samples the LIME map to the common H × W grid used by SHAP and Grad-CAM prior to fusion, as specified in Section 3.3.

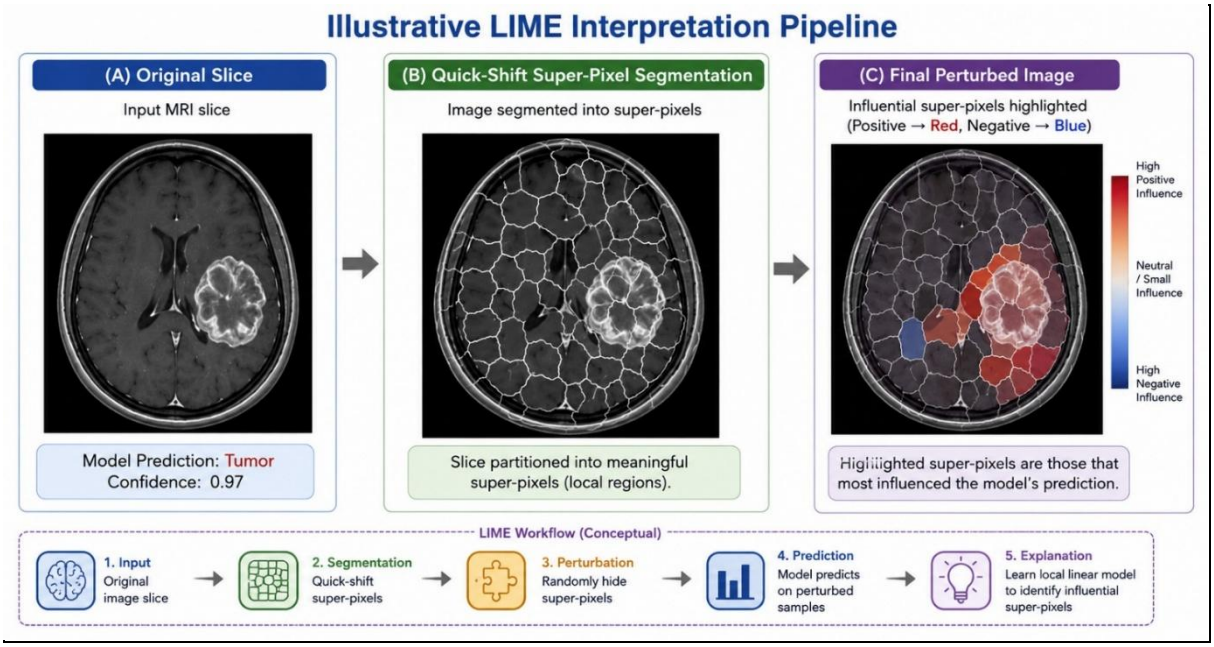


Fig. 4. Illustrative LIME interpretation pipeline: (A) original slice, (B) quick-shift super-pixel segmentation, (C) final perturbed image highlighting influential super-pixel regions. Reproduced from Gaur et al. [3].

5.3 Expected Fusion Behavior

Applying the consensus operator $C(x) = \alpha\tilde{S} + \beta\tilde{L} + \gamma\tilde{G}$ to the maps illustrated in Figures 2–4 with $\alpha = \beta = \gamma = 1/3$ is expected to produce a composite map in which the lesion-adjacent region – scored highly by SHAP, LIME, and Grad-CAM alike, despite their differing spatial granularity – receives reinforced weight, while regions flagged by only one technique (for instance, isolated high-variance SHAP pixels far from the visible lesion, of the kind visible as scattered red/blue speckle in Figure 2) are proportionally suppressed. The associated agreement score $A(x)$ is expected to be highest in exactly this lesion-adjacent region and lowest in anatomically implausible areas where the three techniques disagree, giving the consensus region $R(x)$ a direct interpretation as the sub-area of the composite map that a clinician can trust with the least additional scrutiny.

6. Discussion: Expected Advantages and Computational Cost

Relative to the binary ensemble baseline [1], TXCE is expected to retain comparable robustness benefits from multi-backbone aggregation while extending coverage to the clinically more informative four-class subtype task, at the cost of requiring per-backbone macro-F1 estimates to compute ensemble weights w_m rather than treating all backbones as equally reliable. Relative to the single-backbone multiclass baseline [2], TXCE trades a single, computationally cheap forward pass for three backbone forward passes, an overhead consistent with the inference cost already accepted by the binary ensemble baseline, in exchange for the robustness benefits of architectural diversity and a fused, quantitatively self-checked explanation rather than a single, unverified SHAP map. Relative to the dual-input baseline [3], TXCE incorporates the same noise-robust branch as one ensemble member rather than as the sole classifier, which is expected to recover the higher single-split accuracy achieved by the other two backbones while preserving the dual-input branch's specific robustness advantage on low-quality inputs.

OPEN ACCESS © CINEFORUM

The added cost of the Consensus Explanation Module itself is bounded: because CEM operates only on the ensemble's winning class $\hat{y}$ rather than on all K candidate classes, its cost is one SHAP computation, one LIME computation, and one Grad-CAM computation per explained slice – identical to the per-slice explanation cost already incurred by the binary ensemble baseline that uses all three techniques [1], and substantially cheaper than the K SHAP computations required per slice by the single-technique multiclass baseline [2] once $K > 1$. The fusion, normalization, and agreement-score computation in Section 3.3 involve only elementwise array operations over an $H \times W$ map and are computationally negligible relative to the three underlying explanation computations.

7. Limitations and Threats to Validity

This paper's primary contribution is the TXCE architecture, its mathematical formulation, and its algorithmic specification, grounded in a structured design rationale drawn from results reported for its constituent components in the literature. It does not report new end-to-end empirical training or testing of the fully assembled ensemble and Consensus Explanation Module, and the accuracy, robustness, and fidelity figures TXCE is expected to achieve should be understood as design targets motivated by, rather than guaranteed to match or exceed, the baseline figures in Table 1. Full empirical validation – training all three backbones on the proposed combined corpus, computing ensemble weights, and measuring both classification accuracy and deletion/insertion explanation fidelity – is necessary before any claim of superiority over the baseline systems can be substantiated, and constitutes the immediate next step for this work.

A second limitation concerns the fusion weights α, β, and γ in the Consensus Explanation Module: while Section 4.3 proposes tuning these weights against a deletion/insertion fidelity criterion, no such tuning has yet been performed, and the uniform 1/3-weighting used in the illustrative discussion of Section 5.3 is a design default rather than an empirically optimized configuration. A third limitation is that, like all three baseline systems it builds on, TXCE's proposed evaluation protocol relies on a single-institution, single-scanner data source; external, multi-institution validation is required before any deployment claim, edge-latency figure, or robustness claim can be considered clinically established. Finally, the relabeling procedure proposed in Section 4.1 for augmenting the four-class corpus with Br35H's binary-labeled slices introduces a dependency on the ensemble's own predictions and should be treated as a data-augmentation heuristic requiring expert adjudication rather than an independent source of ground truth.

8. Conclusion and Future Work

This paper has proposed TXCE, a unified architecture for multiclass brain tumor MRI classification that combines a performance-weighted, architecturally diverse backbone ensemble with a novel Consensus Explanation Module fusing SHAP, LIME, and Grad-CAM into a single composite saliency map and a quantitative cross-method agreement score. We formalized the weighted ensemble aggregation rule and the consensus fusion operator mathematically, specified a complete training, evaluation, and edge-cloud deployment protocol, and grounded the architecture's design choices in accuracy, robustness, and computational-cost figures reported for its constituent components in prior literature, supported by illustrative case studies of the underlying explanation techniques. The immediate direction for future work is full empirical realization of this specification: training all three backbones on the proposed combined MRI corpus, computing the performance-weighted ensemble, tuning the consensus fusion weights against a deletion/insertion fidelity criterion, and evaluating the resulting system – including a structured radiologist user study of the consensus region $R(x)$ – against the classification-accuracy and explanation-fidelity baselines established in this paper, with external, multi-institution data as a subsequent validation step toward clinical deployment.

References

1. A. Amin, K. Hasan, S. Zein-Sabatto, D. Chimba, I. Ahmed, and T. Islam, "An explainable AI framework for artificial intelligence of medical things," arXiv preprint arXiv:2403.04130, 2024.
2. S. Ahmed, S. N. Nobel, and O. Ullah, "An effective deep CNN model for multiclass brain tumor detection using MRI images and SHAP explainability," in Proc. 2023 Int. Conf. on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2023.

OPEN ACCESS © CINEFORUM

3. L. Gaur, M. Bhandari, T. Razdan, S. Mallik, and Z. Zhao, "Explanation-driven deep learning model for prediction of brain tumour status using MRI image data," Frontiers in Genetics, vol. 13, p. 822666, 2022.
4. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.
5. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.
6. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. on Computer Vision, pp. 618–626, 2017.
7. V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in Proc. British Machine Vision Conf. (BMVC), 2018.
8. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," IEEE Trans. Neural Networks and Learning Systems, vol. 28, no. 11, pp. 2660–2673, 2017.
9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
10. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 770–778, 2016.
11. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4700–4708, 2017.
12. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2818–2826, 2016.
13. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. on Machine Learning (ICML), pp. 6105–6114, 2019.
14. S. Bhuvaji, "Brain tumor classification (MRI)," Kaggle Dataset, 2020. doi:10.34740/kaggle/dsv/1183165.
15. A. Hamada, "Br35H: Brain tumor detection 2020," Kaggle Dataset, 2021.
16. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.
17. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226–239, 1998

OPEN ACCESS © CINEFORUM