Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable, Uncertainty-Aware Ensemble Learning for Brain Tumor Classification from MRI with a Simulated Fairness Audit

Md. Sajjad Ullah
University of Asia Pacific (UAP)
Shochi Akter
American International University- Bangladesh (AIUB)
Sadman Sadik Khan
sadman15-13696@diu.edu.bd

Daffodil International University

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11445762/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable, Uncertainty-Aware Ensemble Learning for Brain Tumor Classification from MRI with a Simulated Fairness Audit

Md. Sajjad Ullah[1], Shochi Akter[2] & Sadman Sadik Khan[3*]

[1]Department of Computer Science and Engineering, University of Asia Pacific (UAP), Dhaka, Bangladesh.
[2]Department of Computer Science and Engineering, American International University-Bangladesh (AIUB), Dhaka, Bangladesh.
[3]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh.
*email: sadman15-13696@diu.edu.bd

Abstract

Most brain MRI classifiers stop at accuracy; whether their explanations hold up, how sure each call is and whether some patients fare worse is rarely checked. We combined five convolutional networks (Xception, InceptionV3, EfficientNetB4, DenseNet121 and ConvNeXtTiny) and a ViT-B/16 transformer through a logistic regression meta-learner fitted on Out-of-Fold predictions. The models were trained with five-fold cross-validation on 5,040 of 7,200 class-balanced MRI images and tested once on 1,080 held-out images. The MetaEnsemble reached 98.33% accuracy, an AUC-ROC of 0.9989 and an Expected Calibration Error of 0.0083, whereas every single model was underconfident. We set LIME, SHAP, Grad-CAM and Grad-CAM++ side by side and scored the Grad-CAM maps (localization accuracy, Pointing Game accuracy, AOPC faithfulness) instead of judging them by eye. Forty Monte Carlo Dropout passes per image gave each case an uncertainty estimate. Lacking patient metadata, we simulated age and sex labels from epidemiological priors to demonstrate a fairness audit; the gaps were small but pointed the same way in every model (age 1.55% to 4.00%, sex 1.07% to 2.05%). The evidence comes from one public dataset, but it shows that explanation scoring, uncertainty estimation and subgroup checks can be planned into a brain tumor classifier from the start.

Introduction

More than 300,000 people worldwide are diagnosed with a brain or central nervous system cancer every year[1], and which tumor type a patient develops depends partly on age and sex[2]. MRI is the standard way to find gliomas, meningiomas and pituitary tumors without surgery[3,4]. Reading those scans well is a skill built over years of subspecialty training, and even trained readers split on borderline cases; in hospitals without a neuroradiologist on staff, the bottleneck is worse. Deep networks, first CNNs and now Vision Transformers, have come within reach of expert scores on public benchmarks. The review literature has not been won over. It wants subgroup performance screened as a matter of routine[5], explainability, uncertainty and fairness tested side by side[6], and the demographic bias gap that keeps appearing across studies closed[7,8]. We set ourselves a smaller target: one pipeline in which interpretability, predictive uncertainty and a fairness audit are planned from the first experiment, not attached to a finished classifier.

Accuracy is no longer the bottleneck. CNN ensembles built on pretrained backbones[9–12], transformer and hybrid designs[13–17], and a separate stream of work on explainability, uncertainty and fairness[18–30]

have taken care of that. The related work below goes through them one by one, and the same omission turns up in nearly all of them.

That pattern can be stated as three gaps. Explainability methods are almost always judged by eye, with no numerical score for localization or faithfulness. Monte Carlo Dropout and similar uncertainty methods appear mostly in segmentation work or with a single backbone, and they are seldom connected to calibration or to out-of-distribution inputs. The third gap is fairness: the reviews[5–8] keep asking for audits across sex and age, yet we could not find a primary study that runs one. This paper takes on the three gaps together. The ensemble contains five convolutional networks, Xception[31], InceptionV3[32], EfficientNetB4[33], DenseNet121[34] and ConvNeXtTiny[35], plus a ViT-B/16 transformer[36]. Each is trained with five-fold cross-validation on the training split of a balanced four-class MRI dataset of 7,200 images[37], and a logistic regression meta-learner fitted on Out-of-Fold predictions combines them[38]. We then compare four explainability methods, Grad-CAM[39], Grad-CAM++[40], SHAP[41] and LIME[42], and score the Grad-CAM maps with localization accuracy, Pointing Game accuracy[43] and AOPC faithfulness[44]. Predictive uncertainty comes from Monte Carlo Dropout[45,46]. Calibration[47] and out-of-distribution detection[48,49] are measured as well, and fairness gaps across sex and age are estimated with simulated demographic labels. The MetaEnsemble reaches 98.33% accuracy and an AUC-ROC of 0.9989, but we see the method as the main contribution. As far as we know, and among the studies reviewed below, no earlier brain tumor classification study has put numerically scored explanations, calibrated uncertainty and a demographic fairness audit into the same pipeline. Figure 1 outlines the full pipeline. Our contributions are:

- **Multi-model ensemble with a Vision Transformer.** Five CNNs and ViT-B/16 are merged by an Out-of-Fold meta-learner, which keeps test data out of the stacking step. The ensemble reaches 98.33% accuracy and an AUC-ROC of 0.9989 on four-class brain tumor MRI.
- **Multi-method explainability with quantitative scoring.** We place LIME, SHAP, Grad-CAM and Grad-CAM++ side by side, and we score the Grad-CAM maps of all six models with localization accuracy, Pointing Game accuracy and AOPC faithfulness.
- **Monte Carlo Dropout uncertainty quantification.** For every image, forty stochastic forward passes yield a mean confidence and a predictive entropy, so that doubtful cases can be passed to an expert.
- **Calibration and out-of-distribution analysis.** Reliability diagrams, Expected Calibration Error and Brier scores show whether a stated 90% really means 90%, while Maximum Softmax Probability and Energy scores check what the models do when given noise instead of a brain scan.
- **Demographic fairness evaluation.** Accuracy, macro-F1 and AUC are broken down by biological sex and by four age bands for each of the six architectures. The labels behind this breakdown are simulated, so what we report is a worked example of the audit, not a measurement of bias.

Related work

The related work is organized by theme, five in all: CNN transfer learning, transformers and ensembles that mix architectures, explainable AI, uncertainty quantification and demographic fairness. For each

we note the accuracy reported, what it cost to get there and what was left out, and the section closes on the one omission all five share. Nearly every study cited here dates from 2020 to 2026, and Table 1 summarizes the most representative of them.

Most brain tumor classifiers still begin with a CNN pretrained on ImageNet and fine-tune it. Hossain et al.[9] tested six architectures (VGG16, VGG19, InceptionV3, ResNet50, InceptionResNetV2 and Xception), joined the best three into the IVX16 ensemble and explained its decisions with Layer-wise Relevance Propagation and LIME. Their results made a good case for ensembles with XAI. They did not, however, report uncertainty, calibration or subgroup performance. Shah et al.[10] went the other way and fine-tuned a single EfficientNet with heavy augmentation, checking its attention with Grad-CAM. One network is cheaper to train and run than the setup in ref. 9; the price is the extra accuracy that only fusing backbones seems to deliver. Vimala et al.[11] worked through EfficientNet B0 to B4 and settled on B2 as the best value for compute, but they classified only three tumor types. Mathivanan et al.[12] repeated the now familiar finding that pretrained backbones beat training from scratch. It confirms refs. 9–11 but proposes no new model. Seen side by side, ref. 9 pays in speed for ensemble accuracy, ref. 10 pays in diversity for efficiency, and ref. 11 narrows the class list in order to search more architectures. Calibration, predictive uncertainty and subgroup results are missing from all four.

Transformer papers arrived a little later. LCDEiT, from Ferdous et al.[13], is a data-efficient image transformer whose attention scales linearly; it keeps pace with a standard ViT on accuracy while costing much less to run. It is, however, a single model. Afzal and Lee[14] took EfficientNet-B4, added channel and spatial attention plus Monte Carlo Dropout (UA-EffNet-DA), and reached 98.73% and 99.23% accuracy on two benchmarks. Because it reports uncertainty, their work is closer to clinical needs than ref. 13, but calibration, out-of-distribution behavior and fairness go untested, and again there is only one backbone. Only three groups have gone as far as true ensembles of different backbones. Singh et al.[15] combined a custom CNN, ResNet-50 and EfficientNetB5 with adaptive weights and explained the result with five XAI techniques. BRAIN-META, by Napa et al.[16], is far bigger: ten hybrid CNN and ViT backbones feed stacked logistic regression and XGBoost meta-learners, which multiplies the training cost for only a small gain in diversity. Nahiduzzaman et al.[17] kept things light with a hybrid of CNN features and classical machine learning, explained by SHAP, but offered no spatial maps at all. These designs are close to ours. Still, none of refs. 13–17 runs a formal subgroup analysis or applies Monte Carlo Dropout across different architectures, so the gap we saw in the transfer learning studies remains even as the ensembles grow.

Most authors agree that clinicians will not adopt a model they cannot interrogate. Musthafa et al.[18] fine-tuned ResNet-50 on all four classes of the same Kaggle dataset we use and inspected it with Grad-CAM, which makes their work a direct point of comparison; the attribution maps looked anatomically sensible, but only one model and one method were examined. Nazir et al.[19] applied SHAP, LIME and Grad-CAM together. On their binary detection task, the combination was more informative than any method alone, although a binary problem is easier than the four-class one in ref. 18. Akgundogdu and Celikbas[20] saw something similar on BraTS 2019: Grad-CAM gave wide, blurry regions and LIME gave small, tight ones. Used together they cover each other's weaknesses, which is what ref. 19 had found too. Ahmed et al.[21] paired a Vision Transformer with a GRU on hospital data from Bangladesh and reported attention maps, SHAP and LIME under tenfold cross-validation. Of refs. 18–20, theirs is closest to our transformer-based design, and the patient population is similar, but again a single backbone is used. Iftikhar et al.[22] argued instead for deliberately simple CNNs that trade a little accuracy

for transparency. Across this group the methods become more elaborate, from one explainer on one model to several explainers and transformer attention, with ref. 22 as the exception. What none of refs. 18–22 does is put a number on explanation quality. Every one of them relies on looking at heatmaps, and that gives no fair way to compare methods or architectures.

Read side by side, the reviews cited above come to one conclusion from four directions. Dorfner et al.[5] point out that a model can encode age and sex even when neither is an input, and they recommend screening subgroups before any deployment. Solanki et al.[6], surveying detection and classification methods, list explainability, uncertainty and demographic fairness among the field's open problems. Abraham et al.[7] reviewed more than 60 XAI studies on brain tumors and describe demographic bias as critical and underexplored, and Hamza and Damasevicius[8] reach the same view in a PRISMA review of 49 deep learning studies. Different as their methods are, refs. 5–8 leave the same questions open, and those are the questions this paper picks up.

Most uncertainty estimation in medical imaging has been done for segmentation, not classification. The QU-BraTS benchmark of Mehta et al.[23], run inside the MICCAI BraTS 2020 challenge, pitted Monte Carlo Dropout against deep ensembles and Bayesian neural networks. Dropout came out cheapest, and with six backbones to sample, that mattered to us. Kendall and Gal[24] split uncertainty into an aleatoric part (noise in the data) and an epistemic part (gaps in the model) and found that modeling both made predictions more reliable. That split says more than the single number in ref. 23, at a higher computational price. Zou et al.[25], reviewing the area, point out how little of it concerns classification, which is a good reason to carry the benchmark of ref. 23 over to classification. Lakshminarayanan et al.[26] got well calibrated uncertainty from plain ensembles, with no Bayesian machinery required. It is easier to implement than ref. 24 but cannot tell the two kinds of uncertainty apart. Efficiency is the priority in refs. 23 and 26; detail is the priority in ref. 24. Whichever route is taken, ref. 25 shows it has seldom left segmentation. Running MC Dropout over a diverse classification ensemble, reporting it for each class, and tying it to calibration and out-of-distribution behavior in one pipeline is still open.

Table 1. Comparative summary of representative prior studies.

Ref.	Approach	Explainability	Uncertainty	Fairness Audit	Key Gap vs. This Study
9	Hossain et al.: IVX16, 3-model CNN ensemble (InceptionV3, VGG16, Xception)	LRP + LIME	No	No	No calibration, uncertainty or subgroup analysis
10	Shah et al.: single fine-tuned EfficientNet	Grad-CAM	No	No	Single backbone; no uncertainty or fairness
11	Vimala et al.: EfficientNet B0–B4 comparison	—	No	No	Only 3 tumor classes; no XAI/uncertainty/fairness
14	Afzal & Lee: EfficientNet-B4 + attention (UA-EffNet-DA), 98.73%/99.23% acc.	—	MC Dropout	No	Single backbone; no calibration or fairness
15	Singh et al.: dynamic CNN + ResNet-50 + EfficientNetB5 ensemble	5 XAI methods (qualitative)	No	No	No numerical XAI validation, uncertainty or fairness
16	Napa et al.: BRAIN-META, 10 hybrid CNN–ViT backbones	—	No	No	High compute cost; no uncertainty or fairness
17	Nahiduzzaman et al.: lightweight hybrid CNN-ML	SHAP only	No	No	No spatial visual explanation; no fairness
18	Musthafa et al.: single ResNet-50 on same Kaggle corpus	Grad-CAM	No	No	Single architecture and single XAI method
19	Nazir et al.: binary detection benchmark	SHAP + LIME + Grad-CAM (qualitative)	No	No	Binary task only; no quantitative XAI metrics
21	Ahmed et al.: ViT + GRU, primary hospital dataset (Bangladesh)	Attention + SHAP + LIME	No	No	Single backbone; no calibration or fairness

This study	**6-model CNN + ViT ensemble, OOF meta-learner, 98.33% acc., AUC 0.9989**	**LIME + SHAP + Grad-CAM + Grad-CAM++; Grad-CAM scored quantitatively**	**MC Dropout; ensemble ECE = 0.0083**	**Simulated: age & sex, 6 architectures**	—

The comparison covers only the methodological scope reported in each cited study. "—" means not applicable.

The fairness literature in medical imaging is richest outside neuro-oncology. Ricci Lara et al.[27] found performance gaps across age, sex and ethnicity to be common and proposed equity gap metrics to measure them; we use these metrics here. Seyyed-Kalantari et al.[28] showed that leading chest radiograph classifiers underdiagnose underserved groups, a narrower result than ref. 27 but one framed in terms of patient outcomes. Gichoya et al.[29] found that deep models can recover protected attributes from the pixels alone, even in heavily degraded images, which helps explain how disparities like those in ref. 28 slip past unnoticed. In Larrazabal et al.[30], skewing the sex ratio of the training set was all it took to bias a classifier. Where ref. 29 locates the problem in the model, ref. 30 locates it in the data. Between them, refs. 27–30 look at metrics, patient outcomes, model behavior and data composition, and all four conclude that bias in imaging models is real, can be measured and has several sources. Nobody has yet done this kind of analysis for brain tumor classification, where performance by subgroup is still largely unreported.

Put the five themes next to each other and one shortcoming survives every change of architecture, explainer, uncertainty method or fairness metric. The ensemble papers chase accuracy and drop uncertainty and fairness along the way. The XAI papers improve their heatmaps but never put a score on them. The uncertainty papers measure confidence carefully, mostly for segmentation or for one backbone at a time. The fairness papers are rigorous but have not looked at brain tumors. The four reviews see the same hole from four sides. What is missing is a study that measures interpretability, uncertainty and fairness together, and with numbers. Ours tries to be that study. Alongside accuracy, we put three questions to a six-model ensemble. Can its explanations be scored? Do its confidence values mean what they say? Does it work as well for one patient group as for another?

Results

Training behavior

Across the five folds the six architectures trained in much the same way; Table 2 lists what happened in each phase. In this section "validation" refers to the held-out fold of each split, not to the separate validation split of Table 3.

Table 2. Training behavior across the two phases.

Aspect	Phase 1: head warm-up (5 epochs)	Phase 2: full fine-tuning (≤ 55 epochs)
Trainable parameters	Classification head only; backbone frozen	Backbone and head optimized jointly
Training accuracy	Improves from roughly 50% to 70%	Rises rapidly, then remains below validation accuracy (Xception: 0.8269 at epoch 55)
Validation accuracy	Comparatively low	Converges towards the training curve by epoch 15–25 (Xception: 0.9752 at epoch 55)
Focal loss	—	0.5–0.7 at epoch 1, falling to 0.18–0.27 by the final epoch
Validation AUC	—	Steepest gains between epochs 15 and 35; diminishing returns thereafter

Aspect	Phase 1: head warm-up (5 epochs)	Phase 2: full fine-tuning (≤ 55 epochs)
Early stopping	Not applied	Patience 12 on the held-out fold AUC; some InceptionV3 and DenseNet121 folds stopped early, several Xception folds ran to epoch 55
Checkpoint retained	—	Best EMA checkpoint per fold, selected on held-out fold AUC

Validation accuracy never fell below training accuracy during fine-tuning. That is expected given our augmentation: MixUp, CutMix, RandAugment and random erasing make the training task much harder than the clean validation task, which is the reverse of what overfitting looks like. The EMA weights lagged the raw weights a little, but their loss curves were smoother and their validation AUC higher. Table 4 sets each model's mean cross-validation accuracy beside its test accuracy; the largest difference is 0.69 percentage points, for EfficientNetB4.

Table 4. Validation and test accuracy across the six architectures.

Model	Validation Accuracy	Test Accuracy
Xception	97.30%	96.85%
InceptionV3	97.38%	97.41%
EfficientNetB4	94.86%	94.17%
DenseNet121	90.02%	89.91%
ConvNeXtTiny	98.12%	97.78%
ViT-B/16	97.88%	97.87%

Validation accuracy is the mean accuracy over the five cross-validation folds.

Test-set performance and ensemble results

We evaluated the held-out test set once, after all training and meta-learner fitting were finished. Table 5 lists the results with ten-view test-time augmentation.

Table 5. Test-set results under TTA ×10.

Rank	Model	Accuracy	Macro-F1	Precision	Recall	AUC-ROC	AUC Rank
1	**MetaEnsemble**	**98.33%**	**0.9833**	**0.9833**	**0.9833**	**0.9989**	1st
2	AUC Ensemble	98.15%	0.9814	0.9816	0.9815	0.9973	3rd
3	ViT-B/16	97.87%	0.9787	0.9789	0.9787	0.9984	2nd
4	ConvNeXtTiny	97.78%	0.9777	0.9778	0.9778	0.9971	5th
5	InceptionV3	97.41%	0.9740	0.9744	0.9741	0.9973	3rd
6	Xception	96.85%	0.9683	0.9690	0.9685	0.9970	6th
7	EfficientNetB4	94.17%	0.9412	0.9420	0.9417	0.9918	7th
8	DenseNet121	89.91%	0.8970	0.9115	0.8991	0.9899	8th

On every aggregate metric the MetaEnsemble comes first. The order by AUC-ROC is not the same as the order by accuracy, though. Among single models ViT-B/16 leads on both accuracy and AUC

(0.9984), while the AUC Ensemble is more accurate than ViT-B/16 yet ranks below it on AUC. The two measures answer different questions: AUC describes how well cases are ranked over all thresholds, accuracy how well they are classified at one. Per-class ROC curves are shown in Figure 2.

To see whether these differences go beyond sampling noise, Table 6 gives a 95% Wilson confidence interval for each model's test accuracy (n = 1,080), along with a two-proportion z-test against the MetaEnsemble.

Table 6. Ninety-five percent confidence intervals and pairwise significance versus the MetaEnsemble.

Model	Accuracy	95% CI (Wilson)	z (vs. MetaEnsemble)	p-value	Sig.
MetaEnsemble	**98.33%**	**(97.38%, 98.94%)**	—	—	—
AUC Ensemble	98.15%	(97.16%, 98.80%)	0.318	0.750	ns
ViT-B/16	97.87%	(96.82%, 98.58%)	0.783	0.434	ns
ConvNeXtTiny	97.78%	(96.72%, 98.50%)	0.925	0.355	ns
InceptionV3	97.41%	(96.28%, 98.20%)	1.481	0.139	ns
Xception	96.85%	(95.63%, 97.74%)	2.243	0.025	*
EfficientNetB4	94.17%	(92.61%, 95.42%)	5.088	< 0.001	***
DenseNet121	89.91%	(87.97%, 91.57%)	8.317	< 0.001	***

ns = not significant at $\alpha = 0.05$; * $p < 0.05$; *** $p < 0.001$. The z-test compares two proportions without pairing, so it is conservative for a shared test set (see Statistical significance testing in Methods).

With this conservative test, the MetaEnsemble's lead over the next three configurations is not significant: the AUC Ensemble ($p = 0.750$), ViT-B/16 ($p = 0.434$) and ConvNeXtTiny ($p = 0.355$). The confidence intervals of these four overlap widely. The direction of the difference is consistent, however, and it becomes significant against Xception ($p = 0.025$) and highly significant against EfficientNetB4 and DenseNet121 ($p < 0.001$ for both). So the accuracy claim needs qualifying, but the case for the MetaEnsemble does not rest on accuracy alone. It has the highest AUC-ROC (0.9989, against 0.9984 for the next best model), it is close to perfectly calibrated (Table 7), and it has the highest meningioma F1-score (0.9760; Table 8), meningioma being the weakest class for every single backbone. A paired McNemar's test on individual predictions would be more powerful than our unpaired test, and we leave it for future validation.

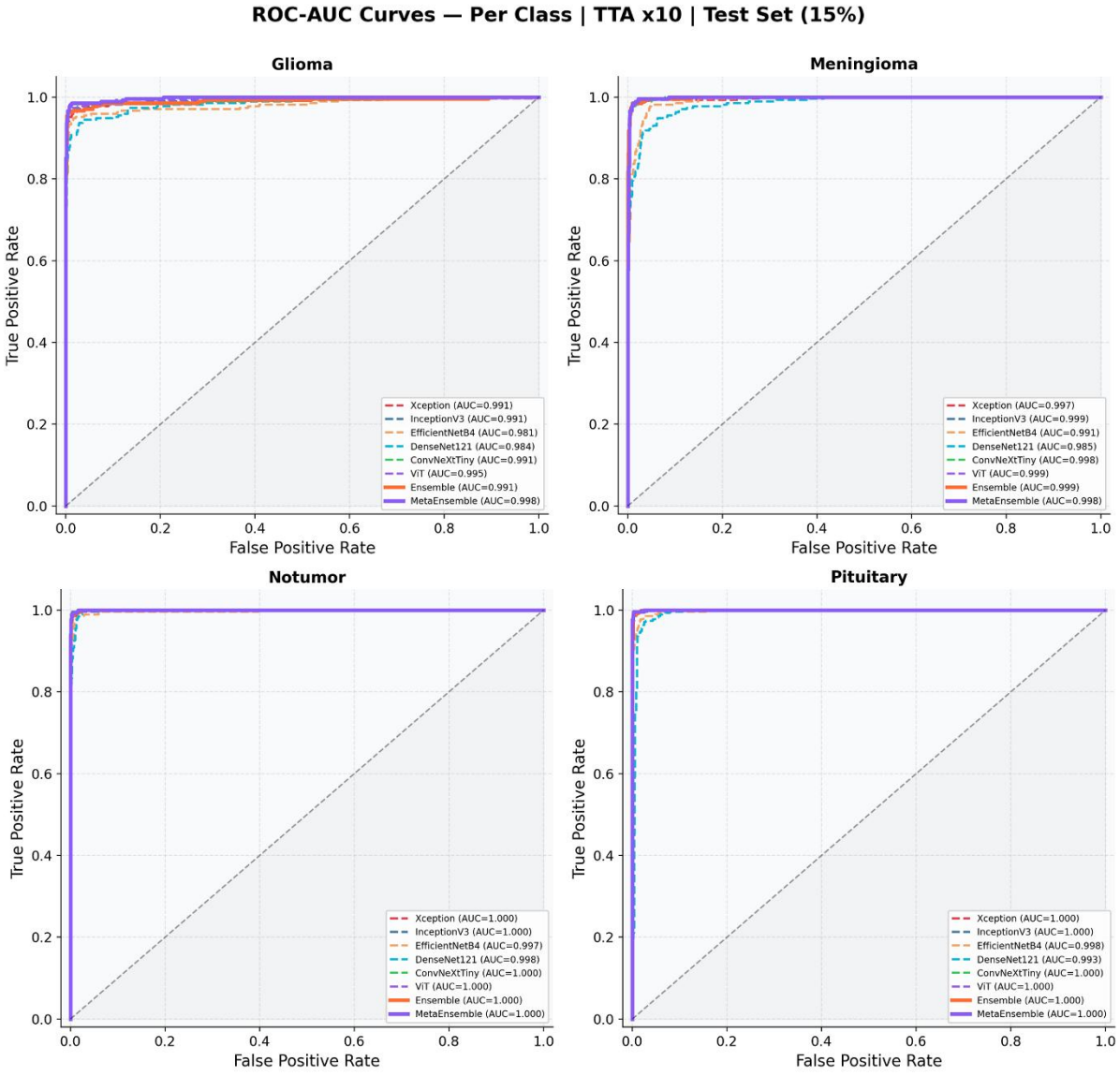


Figure 2. ROC-AUC curves per class across all models and the MetaEnsemble.

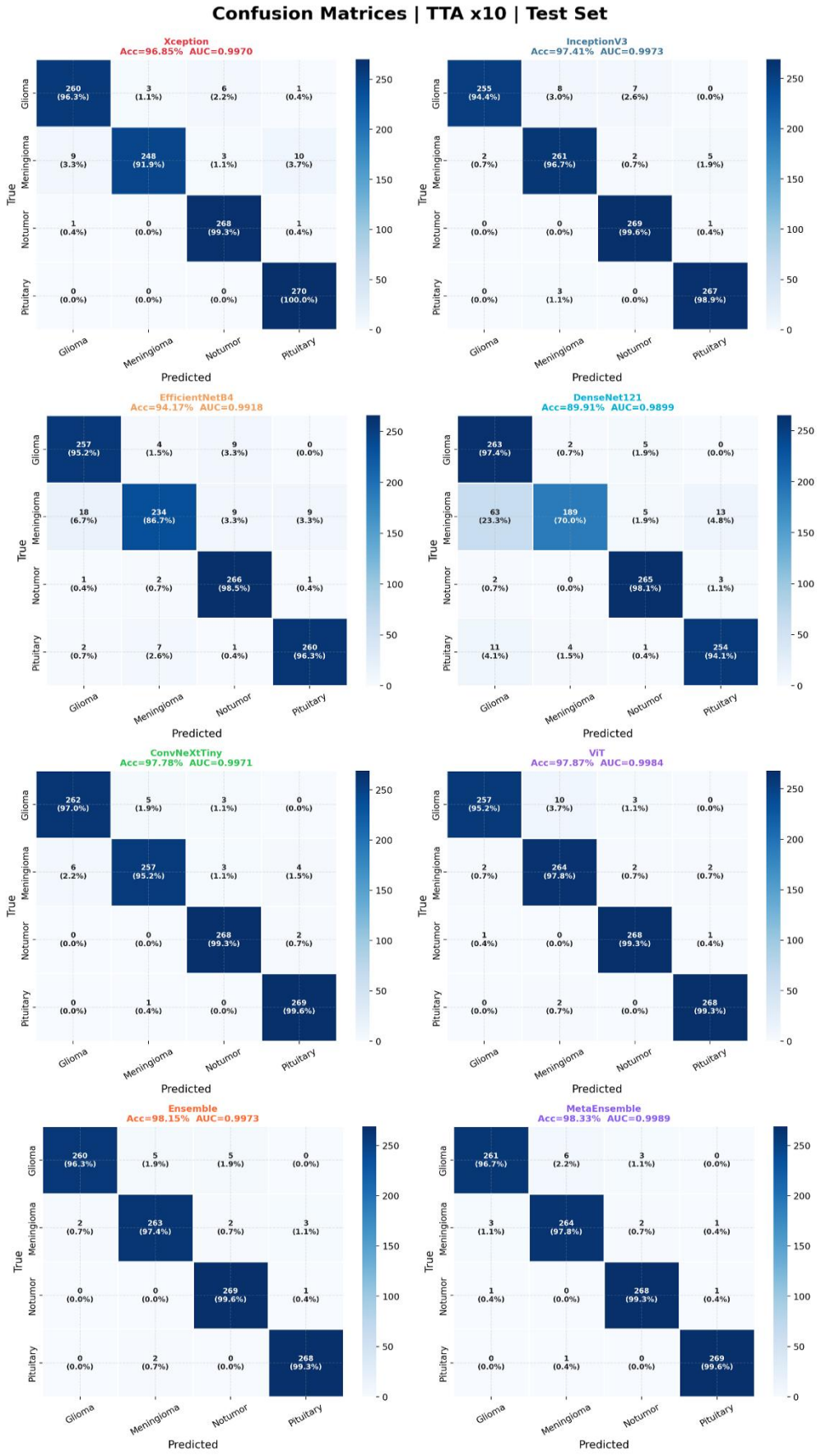


Figure 3. Confusion matrices on the test set.

In every panel the diagonal dominates, and most of the remaining errors sit in the meningioma row. DenseNet121 stands out. It recognizes only 70.0% of meningiomas, and 63 of its 81 meningioma errors are labelled glioma. The MetaEnsemble recognizes 97.8%, the same as the best single model, ViT-B/16. Per-class F1-scores are given in Table 8.

Table 8. Per-class F1-score on the test set.

Model	Glioma	Meningioma	No Tumor	Pituitary
Xception	0.9630	0.9520	0.9799	0.9783
InceptionV3	0.9677	0.9631	0.9818	0.9834
EfficientNetB4	0.9380	0.9052	0.9586	0.9630
DenseNet121	0.8637	0.8129	0.9707	0.9407
ConvNeXtTiny	0.9740	0.9644	0.9853	0.9872
ViT-B/16	0.9698	0.9670	**0.9871**	0.9908
AUC Ensemble	**0.9774**	0.9741	0.9853	0.9889
MetaEnsemble	0.9757	**0.9760**	**0.9871**	**0.9945**

The best value in each column is shown in bold. Recall values for each class are taken from the confusion matrices in Figure 3.

For every single architecture, meningioma is the hardest class, and this is where ensembling pays off most. DenseNet121 shows an instructive asymmetry: its glioma recall is high (97.4%, Figure 3), but its glioma F1 is only 0.8637, because the meningiomas it misclassifies end up in the glioma column and drag glioma precision down. The MetaEnsemble has the highest F1 for meningioma (0.9760) and pituitary (0.9945), and it shares the highest no tumor F1 (0.9871) with ViT-B/16.

Cross-validation performance

Table 9 gives the five-fold cross-validation results on the training split of 5,040 images.

Table 9. Five-fold cross-validation results.

Model	Mean CV Accuracy	Mean CV AUC ± Std
Xception	97.30%	0.9975 ± 0.0007
InceptionV3	97.38%	0.9977 ± 0.0008
EfficientNetB4	94.86%	0.9929 ± 0.0019
DenseNet121	90.02%	0.9869 ± 0.0032
ConvNeXtTiny	**98.12%**	0.9981 ± 0.0010
ViT-B/16	97.88%	**0.9983 ± 0.0007**

The best value in each column is shown in bold. Each value is the mean over the five stratified folds, and the standard deviation shows stability across folds.

ConvNeXtTiny reached the highest mean validation accuracy (98.12%). ViT-B/16 had the highest mean AUC (0.9983) and, together with Xception, the smallest spread across folds (± 0.0007). DenseNet121 was weakest on both counts, with the lowest mean accuracy (90.02%) and the most variable AUC (± 0.0032). We kept it anyway, because its dense skip connections supply features that the meta-learner can still exploit.

Cross-validation and testing disagree slightly on the order (Tables 9 and 5). ConvNeXtTiny led the single models in cross-validation, while ViT-B/16 was more accurate on the test set (97.87% against

97.78%). The swap is small, but it shows why we used cross-validation only to choose models and kept the test split for a single final run.

Ablation study

Table 10 and Figure 4 show the effect of each ensemble component. The starting point is Xception with ten-view TTA (96.85%), and each later row adds one component.

Table 10. Ablation study.

Component	Accuracy	Macro-F1	AUC-ROC	Δ Accuracy	Δ AUC
Xception with Test-Time Augmentation (×10), starting point	96.85%	0.9683	0.9970	—	—
+ 3-Model AUC-Weighted Ensemble (Xception, InceptionV3, DenseNet121)	97.59%	0.9758	0.9972	+0.74%	+0.0002
+ 6-Model AUC-Weighted Ensemble (adds EfficientNetB4, ConvNeXtTiny, ViT-B/16)	98.15%	0.9814	0.9973	+0.56%	+0.0001
+ OOF Meta-Learner (final system)	**98.33%**	**0.9833**	**0.9989**	+0.18%	+0.0016

The final system is shown in bold. Each row adds one component to the row above it. The first row is the starting point, so it has no Δ values.

No step made things worse; each gain adds to the one before.

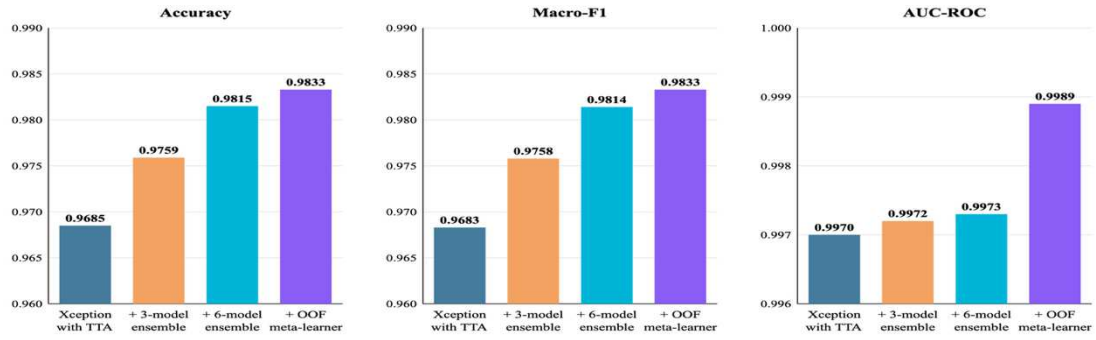


Figure 4. Ablation study.

The accuracy gains shrink as components are added: +0.74% for the three-model ensemble, +0.56% when the other three models join and +0.18% for the meta-learner. AUC follows a different pattern. The averaged ensembles barely change it, and then the meta-learner alone adds 0.0016, which suggests that stacking mostly improves how probabilities are ranked rather than where the decision boundary falls.

Explainability analysis

Qualitative visualization

Full explanation panels were drawn for the subset of 30 images described in Methods. Figure 5 sets heatmaps from different architectures and tumor classes next to each other; Figure 6 is the complete panel for InceptionV3, chosen because it scored best on the Pointing Game and AOPC.

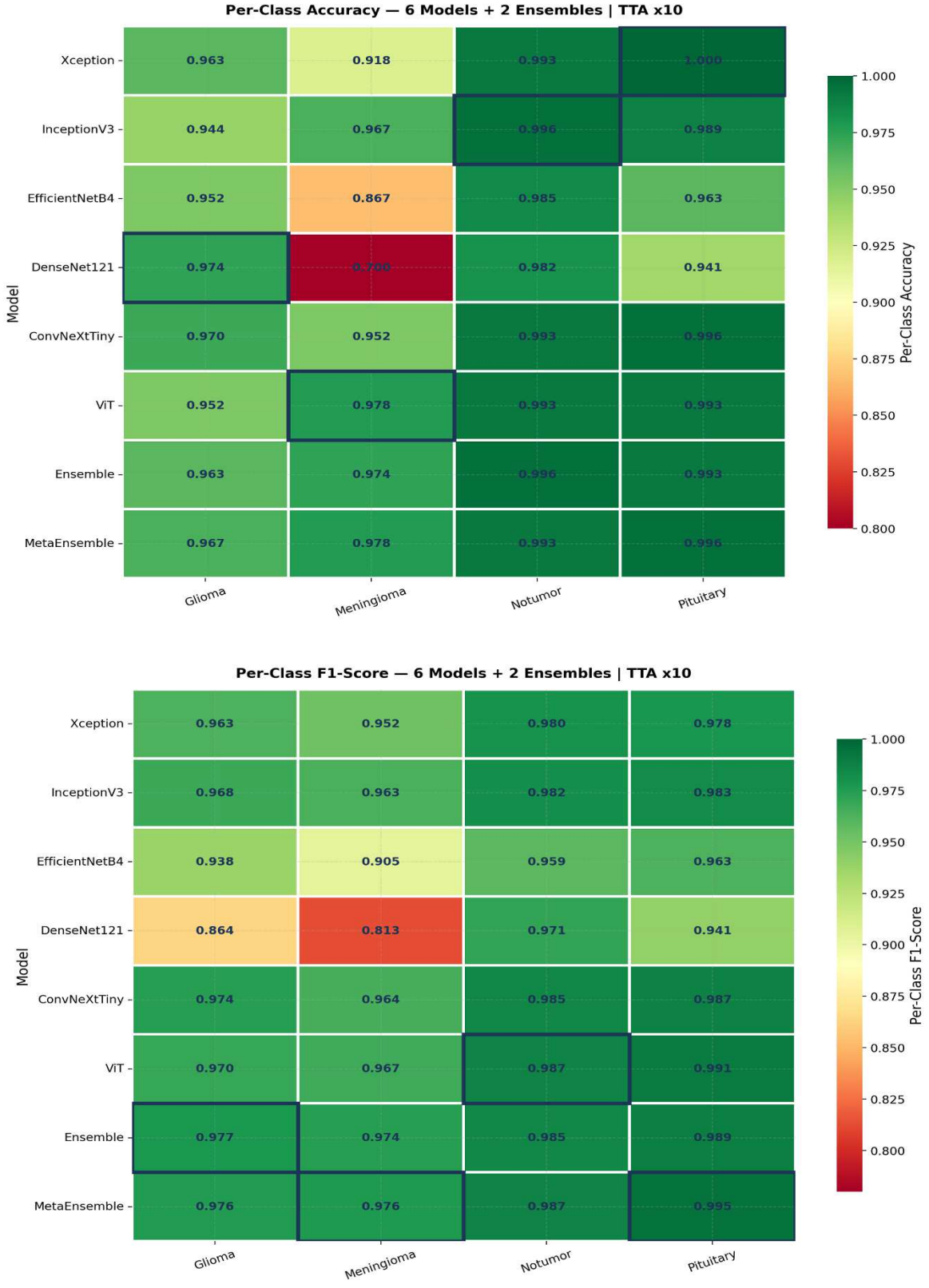


Figure 5. Explanation heatmaps across architectures.

On any one image the four methods agree on where the evidence is and disagree on how crisply to draw it. Grad-CAM, Grad-CAM++, SHAP and LIME work on quite different principles, so their agreement makes it less likely that a highlighted region is an artifact of any one method.

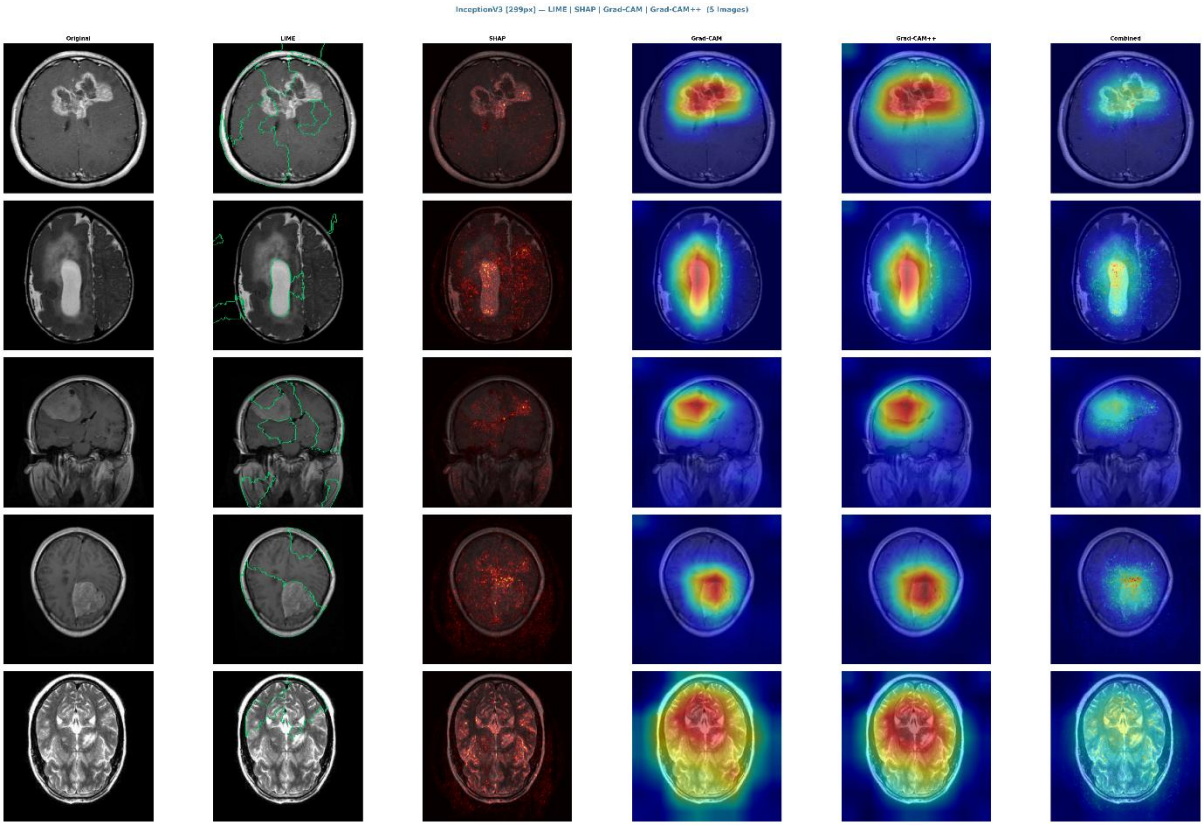


Figure 6. InceptionV3 explainability panel.

Wherever a lesion is visible, all four methods land on it. Grad-CAM++ follows the boundary more closely than Grad-CAM, LIME outlines the same area in superpixels, and the overlay brightens where the gradient and Shapley maps overlap. The bottom row has no focal lesion. There the activation spreads thinly over the whole brain, which is the no tumor pattern listed with the other recurring patterns in Table 11.

Table 11. Qualitative characteristics of the explanation maps.

Aspect	Observation
Glioma	Activation concentrated on the tumor mass and the surrounding perilesional edema
Meningioma	Most variable heatmaps, consistent with the class's lower accuracy and its imaging overlap with adjacent dural structures
Pituitary tumor	Consistent highlighting of the sellar and suprasellar region; Grad-CAM++ noticeably sharper than Grad-CAM
No tumor	Diffuse, low-intensity activation across the parenchyma with no focal hotspots
Grad-CAM vs Grad-CAM++	Grad-CAM++ produces sharper, more localized maps where the relevant structure is small and well defined
LIME	Identifies the same gross anatomical regions as Grad-CAM but with discrete superpixel boundaries
SHAP	Spatially consistent with Grad-CAM, with additional pixel-level granularity along tumor margins

In the qualitative subset of 30 images, LIME failed once for ViT-B/16, and the original image was shown in its place. The other 29 ViT explanations ran without error. This subset is separate from the set of 60 images used in Table 12.

The methods disagree about resolution, not about location.

Quantitative XAI results

Table 12 gives the quantitative XAI results on the set of 60 images defined in Methods.

Table 12. Quantitative XAI evaluation.

Model	Localization Acc. (↑)	Pointing Game (↑)	AOPC Faithfulness (↑)	n
Xception	0.9068	0.9500	0.2728	60
InceptionV3	0.9008	**0.9833**	**0.3579**	60
EfficientNetB4	0.2483	0.0667	0.2772	60
DenseNet121	0.8722	0.8833	0.2330	60
ConvNeXtTiny	**0.9195**	0.9167	0.2240	60
ViT-B/16	0.7144	0.6667	0.1718	60

The best value in each column is shown in bold. A Pointing Game score of 0.9833 equals 59 hits out of 60 images.

InceptionV3 achieved the best Pointing Game score (0.9833, i.e. 59 of 60 images) and the highest AOPC faithfulness (0.3579). Its maps are precise in space and tied to what actually drives its predictions, which makes it the natural model to show clinicians. The highest Localization Accuracy belonged to ConvNeXtTiny (0.9195). EfficientNetB4's poor scores say more about our choice of target layer than about its explanations. Its target, conv_head, is a pointwise 1 × 1 convolution that mixes spatial information and returns an almost uniform gradient. This configuration error can be corrected and is not a property of the model. Grad-CAM has not yet been rerun on a spatially resolved layer such as the last MBConv block, so EfficientNetB4's numbers in Table 12 should not be set against the others. Until that is done, our comparison across models, including the statement about InceptionV3, is limited to the five architectures with a valid spatial target: Xception, InceptionV3, DenseNet121, ConvNeXtTiny and ViT-B/16. ViT-B/16 sits in the middle. Self-attention spreads information over the whole image instead of building local hierarchical features, and its Grad-CAM maps show correspondingly global activation.

Calibration, OOD detection and uncertainty

Table 7 and Figure 7 report probability calibration.

Table 7. Calibration results.

Model	ECE (↓)	Brier (↓)	Test Accuracy	Assessment
MetaEnsemble	**0.0083**	**0.0280**	98.33%	Well calibrated
ViT-B/16	0.2270	0.1049	97.87%	Underconfident
EfficientNetB4	0.2398	0.1740	94.17%	Underconfident
Xception	0.2605	0.1456	96.85%	Underconfident
InceptionV3	0.2967	0.1649	97.41%	Underconfident
AUC Ensemble	0.3227	0.1774	98.15%	Underconfident
ConvNeXtTiny	0.3522	0.2067	97.78%	Underconfident

Model	ECE (↓)	Brier (↓)	Test Accuracy	Assessment
DenseNet121	0.3985	0.3758	89.91%	Underconfident

The best value in each column is shown in bold. Rows are sorted by ascending ECE.

High accuracy is no guarantee of trustworthy confidence. ConvNeXtTiny is fourth on accuracy and seventh on ECE: right most of the time, yet its stated probabilities cannot be taken at face value.

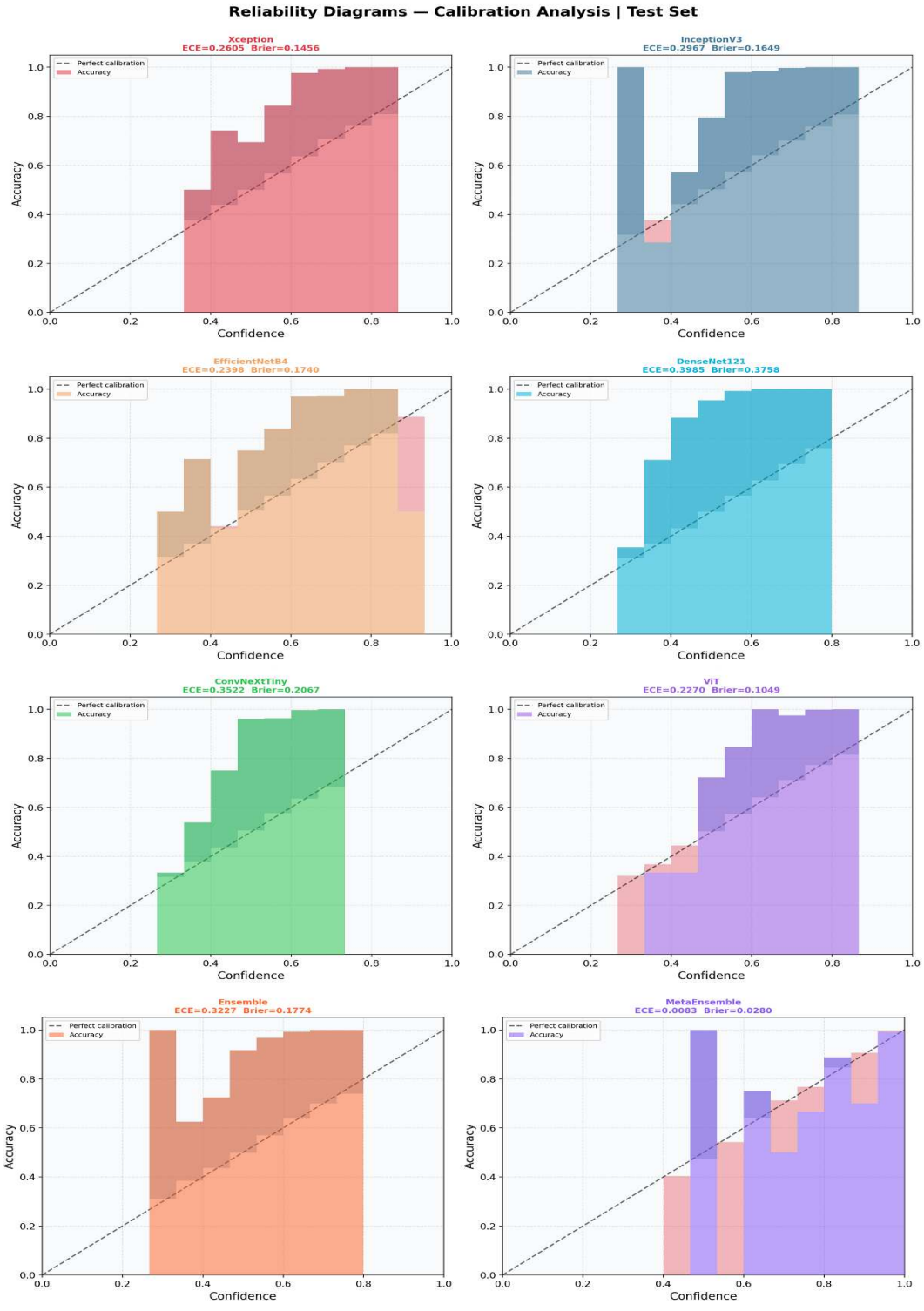


Figure 7. Reliability diagrams.

The MetaEnsemble's bars sit almost on the diagonal, so a prediction made with 90% confidence is correct about 90% of the time. Every single architecture plots above the diagonal; in each bin its accuracy beats its stated confidence. That is underconfidence, and it fits the low mean confidence values in Table 13 and Figure 8. DenseNet121 is furthest off. We suspect the regularization stack, since label smoothing, MixUp, CutMix and Focal Loss each pull probabilities away from 1. The logistic regression meta-learner, fitted on Out-of-Fold probabilities, undoes this without any post hoc temperature scaling.

Table 13 and Figure 8 report out-of-distribution detection and Monte Carlo Dropout uncertainty.

Table 13. Out-of-distribution detection and MC Dropout uncertainty.

Model	MSP AUROC (↑)	MSP AUPR (↑)	MSP FPR95 (↓)	Energy AUROC (↑)	Energy AUPR (↑)	Energy FPR95 (↓)	MC Conf.
ViT-B/16	**1.0000**	**1.0000**	**0.000**	**1.0000**	**1.0000**	**0.000**	**0.642**
ConvNeXtTiny	**1.0000**	**1.0000**	**0.000**	**1.0000**	**1.0000**	**0.000**	0.530
Xception	0.8930	0.9820	1.000	0.9699	0.9788	0.195	0.556
InceptionV3	0.7184	0.9368	0.990	0.6122	0.6492	0.935	0.596
EfficientNetB4	0.4375	0.7630	0.995	0.4843	0.4467	0.540	0.587
DenseNet121	0.1718	0.7043	1.000	0.0000	0.3081	1.000	0.441

The best value in each column is shown in bold. MC Conf. is the mean softmax confidence over 40 stochastic forward passes per test image.

Under both scores, ConvNeXtTiny and ViT-B/16 separate brain MRI from noise perfectly. One explanation would be that layer normalization, GELU and self-attention produce tighter, more bounded feature distributions than the older convolutional designs. Perfect separation from Gaussian and uniform noise is, however, a weak test, and it cannot tell us which architectural feature is responsible. That would need targeted ablations, for example replacing layer normalization with batch normalization in an otherwise unchanged model. We have not run them, so this remains an observation. Xception does well with Energy scoring (AUROC = 0.9699, FPR95 = 0.195) and badly with MSP (FPR95 = 1.000); for all its accuracy, its softmax confidence stays high on inputs it has never seen. DenseNet121's Energy AUROC is 0.0000. That is not chance level but a complete inversion: the model finds every noise image more familiar than every real scan. Its MSP histogram in Figure 8 tells the same story, as does EfficientNetB4's, with a block of noise images at confidence close to 1. With MC Dropout, mean confidence is highest for ViT-B/16 (0.642) and lowest for DenseNet121 (0.441); predictive entropy goes from 0.946 for InceptionV3 up to 1.182 for DenseNet121.

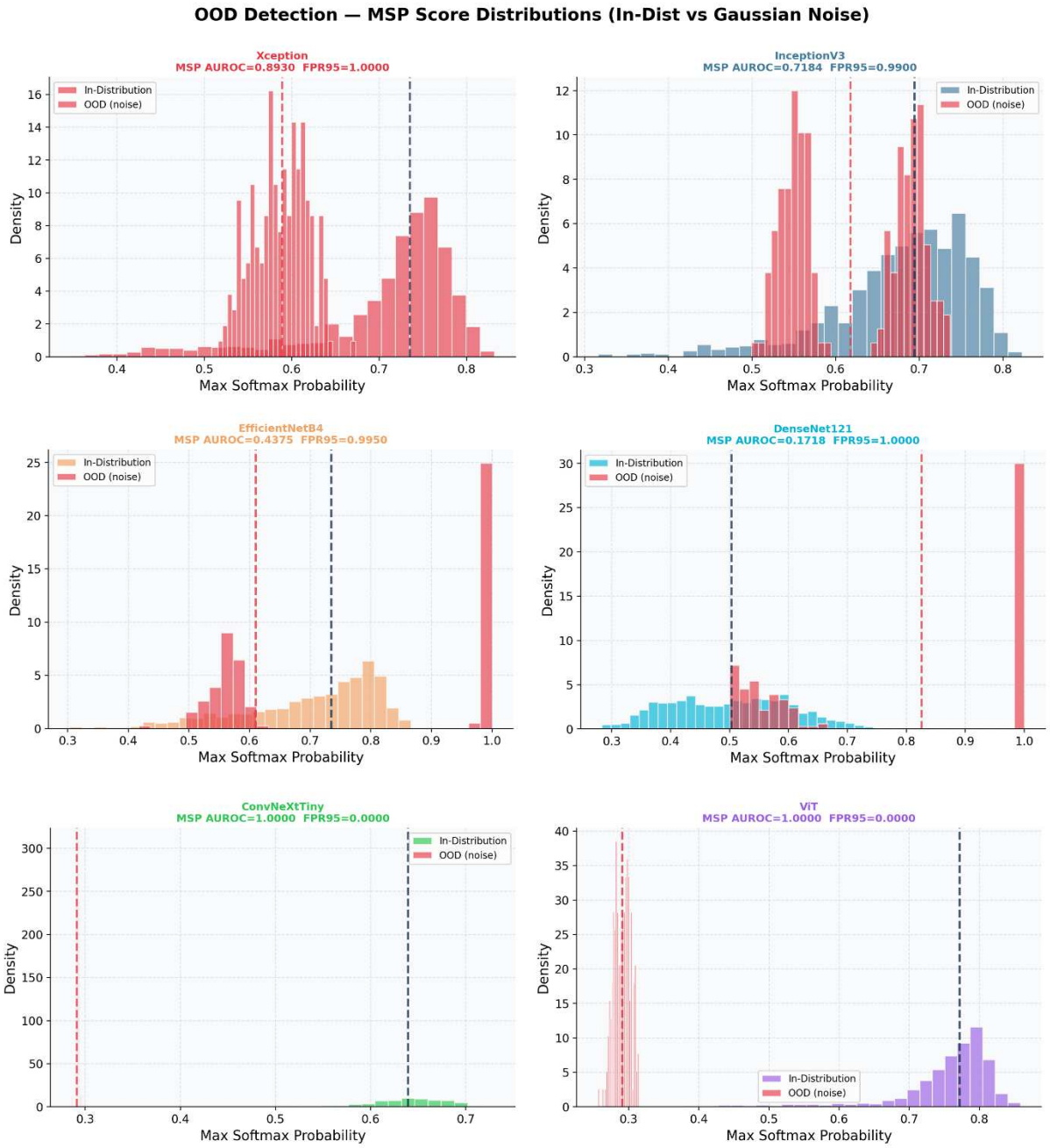


Figure 8. Out-of-distribution MSP score distributions, in-distribution images versus synthetic noise.

Demographic fairness analysis

What follows is a simulated fairness audit. The age and sex labels are synthetic, for reasons given at the end of this section, so the results illustrate an audit method and say nothing definite about real demographic bias in the classifier.

Table 14 and Figures 9 and 10 report accuracy gaps across biological sex and four age groups.

Table 14. Demographic equity gaps.

Model	Sex Acc Gap	Sex F1 Gap	Age Acc Gap	Age F1 Gap	Lowest Subgroup
Xception	**1.07%**	**0.97%**	**1.55%**	**1.49%**	> 70
InceptionV3	1.33%	1.26%	3.19%	3.14%	> 70
EfficientNetB4	1.47%	1.32%	4.00%	4.00%	> 70
DenseNet121	2.05%	1.48%	2.20%	1.68%	> 70
ConvNeXtTiny	1.62%	1.54%	2.53%	2.42%	> 70
ViT-B/16	1.41%	1.31%	1.58%	**1.49%**	> 70

The smallest gap in each column is shown in bold. A gap is the highest minus the lowest subgroup value. Female accuracy is below male accuracy in all six architectures, and patients over 70 have the lowest accuracy in all six.

For every architecture the sex gap stays under 2.1%, whereas age gaps go up to 4.00%. Age is the larger fairness risk.

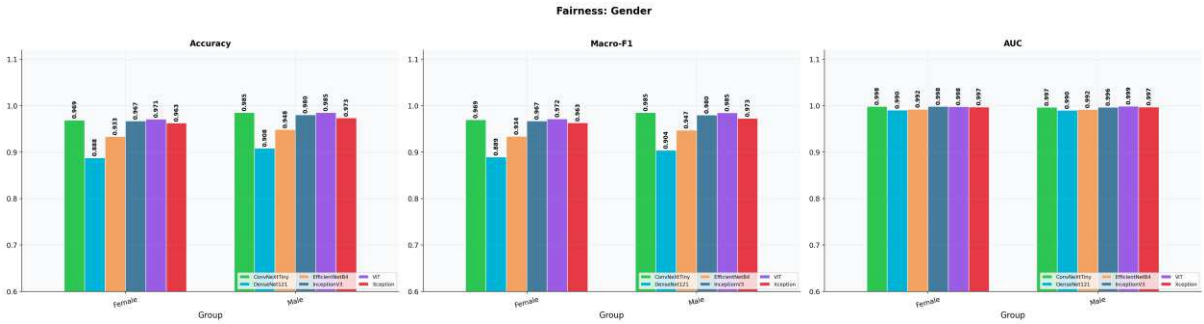


Figure 9. Accuracy, macro-F1 and AUC by biological sex.

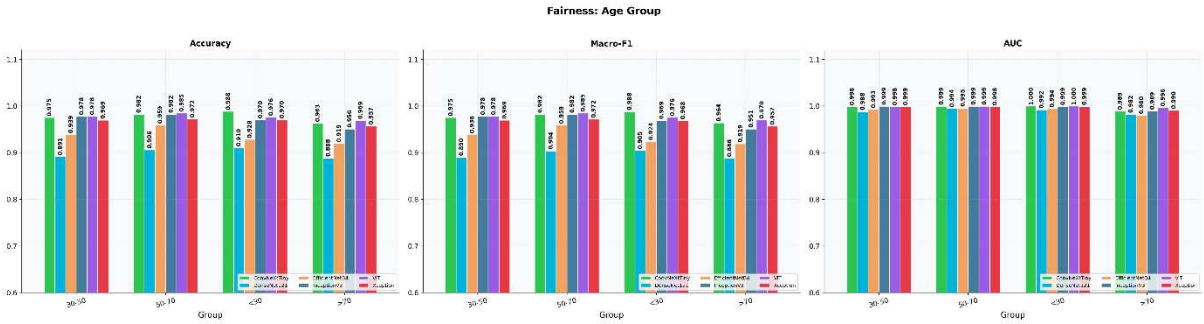


Figure 10. Accuracy, macro-F1 and AUC by age group.

Every architecture loses accuracy in the over-70 group. EfficientNetB4 drops most and has the largest age gap (4.00%); Xception and ViT-B/16 drop least, with bars that are almost level across the four groups. The age labels are simulated, though, so nothing here can be pinned on real patients. If the same drop appeared with real labels, the usual suspects would be structural brain changes with age, comorbidities that change how a scan looks, and too few elderly patients in training.

On both dimensions Xception has the smallest gaps, 1.07% for sex and 1.55% for age. ViT-B/16 is just behind (1.41% and 1.58%) and about one percentage point more accurate, which makes it the better compromise between fairness and performance.

Each subgroup accuracy in Table 14 is a proportion with a known sample size and can be given a 95% Wilson confidence interval. Whether the pattern holds across the six architectures can be examined with a sign test. Table 15 gives the full subgroup breakdown.

Table 15. Subgroup performance with 95% confidence intervals (test set, n = 1,080).

Model	Subgroup	N	Accuracy	95% CI (Wilson)
Xception	Female	481	96.26%	(94.17%, 97.62%)
	Male	599	97.33%	(95.71%, 98.35%)
	Age 30–50	359	96.94%	(94.60%, 98.28%)
	Age 50–70	393	97.20%	(95.06%, 98.43%)
	Age <30	167	97.01%	(93.19%, 98.72%)
	Age >70	161	95.65%	(91.30%, 97.88%)
InceptionV3	Female	481	96.67%	(94.66%, 97.94%)
	Male	599	98.00%	(96.54%, 98.85%)
	Age 30–50	359	97.77%	(95.66%, 98.87%)
	Age 50–70	393	98.22%	(96.37%, 99.14%)
	Age <30	167	97.01%	(93.19%, 98.72%)
	Age >70	161	95.03%	(90.50%, 97.46%)
EfficientNetB4	Female	481	93.35%	(90.76%, 95.25%)
	Male	599	94.82%	(92.74%, 96.33%)
	Age 30–50	359	93.87%	(90.89%, 95.92%)
	Age 50–70	393	95.93%	(93.49%, 97.48%)
	Age <30	167	92.81%	(87.86%, 95.84%)
	Age >70	161	91.93%	(86.68%, 95.22%)
DenseNet121	Female	481	88.77%	(85.64%, 91.29%)
	Male	599	90.82%	(88.24%, 92.88%)
	Age 30–50	359	89.14%	(85.50%, 91.95%)
	Age 50–70	393	90.59%	(87.30%, 93.10%)
	Age <30	167	91.02%	(85.71%, 94.48%)
	Age >70	161	88.82%	(83.02%, 92.81%)
ConvNeXtTiny	Female	481	96.88%	(94.92%, 98.10%)
	Male	599	98.50%	(97.17%, 99.21%)
	Age 30–50	359	97.49%	(95.30%, 98.67%)
	Age 50–70	393	98.22%	(96.37%, 99.14%)
	Age <30	167	98.80%	(95.74%, 99.67%)
	Age >70	161	96.27%	(92.10%, 98.28%)
ViT-B/16	Female	481	97.09%	(95.17%, 98.26%)
	Male	599	98.50%	(97.17%, 99.21%)
	Age 30–50	359	97.77%	(95.66%, 98.87%)
	Age 50–70	393	98.47%	(96.71%, 99.30%)
	Age <30	167	97.60%	(94.00%, 99.06%)
	Age >70	161	96.89%	(92.93%, 98.66%)

N is the number of test images in each subgroup. The sex totals and the age totals each sum to 1,080. Confidence intervals use the Wilson score method.

We ran two nonparametric sign tests on these patterns, and they come with a caveat. In all six architectures, male accuracy is higher than female accuracy. Treated as six independent trials under the

null hypothesis that each model is equally likely to favor either sex, a one-tailed sign test would give $p = 0.5^6 = 0.0156$. But the six models share the same data, labels, preprocessing and test cases. They are not independent, and so this nominal p-value is not a valid test. We give it only as a rough measure of how consistent the models are. Confirming the pattern would require a test that allows for the shared data, such as a cluster-robust or permutation test, which we did not run. Patients over 70 also have the lowest accuracy of the four age groups in all six architectures. The age result invites the same naive arithmetic: with four groups each equally likely to come last, six out of six gives $0.25^6 \approx 2.4 \times 10^{-4}$ (one-tailed sign test), with the same caveat attached. Inside a single model the subgroup intervals often overlap, and the over-70 group has only n = 161 images. What stands out is that the direction of both gaps is the same in every architecture, so the pattern is not the product of one model or one sample. Establishing formal significance would still require a test that accounts for the shared data and test set.

One caveat governs everything in this section. The age and sex labels were simulated from published epidemiological priors, since the source dataset carries no patient metadata. They do not come from patient records, and they are not linked to the people whose scans are in the dataset. This affects the whole fairness analysis. The gaps in Table 14 are not evidence of real demographic bias in our model. The fairness analysis is best seen as a simulated audit protocol that shows the metrics, tests and reporting format a real audit would use, rather than as an assessment of this classifier's bias. Replacing the synthetic labels with real patient demographics is a priority for future work.

Consolidated summary

Table 16 brings all evaluation dimensions together in one comparison.

Table 16. Consolidated multi-dimensional summary.

Model	Accuracy	AUC-ROC	ECE (↓)	MC Conf.	OOD AUROC	Sex Gap	Age Gap	Rank
MetaEnsemble	**98.33%**	**0.9989**	**0.0083**	—	—	—	—	#1
ViT-B/16	97.87%	0.9984	0.2270	**0.642**	**1.0000**	1.41%	1.58%	#2
ConvNeXtTiny	97.78%	0.9971	0.3522	0.530	**1.0000**	1.62%	2.53%	#3
InceptionV3	97.41%	0.9973	0.2967	0.596	0.7184	1.33%	3.19%	#4
Xception	96.85%	0.9970	0.2605	0.556	0.8930	**1.07%**	**1.55%**	#5
EfficientNetB4	94.17%	0.9918	0.2398	0.587	0.4375	1.47%	4.00%	#6
DenseNet121	89.91%	0.9899	0.3985	0.441	0.1718	2.05%	2.20%	#7

The best value in each column is shown in bold. '—' means not applicable at the ensemble level. OOD AUROC uses MSP scoring. Rank follows test accuracy.

The MetaEnsemble leads on accuracy, AUC-ROC and calibration. Among single architectures ViT-B/16 is the strongest all round: best single-model AUC, best calibration, perfect out-of-distribution detection and the second smallest age gap, just behind Xception.

Discussion

We built a six-model ensemble for classifying brain tumors on MRI, in which five convolutional networks and a ViT-B/16 transformer are combined by an Out-of-Fold meta-learner. It reached 98.33% accuracy and an AUC-ROC of 0.9989. Previous work has mostly handled interpretability, predictive uncertainty and demographic fairness as separate and occasional concerns; here they are evaluated together in one pipeline. For explainability we compared LIME, SHAP, Grad-CAM and Grad-CAM++

and scored the Grad-CAM maps with localization accuracy, Pointing Game accuracy and AOPC faithfulness. The MetaEnsemble was well calibrated (ECE = 0.0083), although every individual model was underconfident. Monte Carlo Dropout added an uncertainty estimate for each case, a natural trigger for sending a scan to a radiologist. In the simulated audit, all six architectures showed small age and sex gaps in the same direction. Synthetic labels and a shared test set mean this must be checked again on real patient data. Of the single backbones, ViT-B/16 came out ahead, with the highest AUC, the best calibration, perfect out-of-distribution detection and one of the two smallest age gaps; a dedicated study could test whether transformer features carry a fairness advantage. The broader point is practical. Interpretability, uncertainty estimation and fairness auditing can be planned in from the choice of architecture onward, instead of being added once a model is finished. A high accuracy figure does not by itself make a model fit for clinical use; it also needs explanations that can be measured, confidence that can be trusted and performance that holds across the patients it serves. We hope the protocol used here (scored explanations, Monte Carlo uncertainty and subgroup auditing) will be taken up as a routine evaluation in brain tumor studies, and possibly in other medical imaging tasks.

Limitations

The study has limits that a reader should weigh. Localization was measured against brain masks from Otsu thresholding, not against tumor outlines from radiologists, so it says more about anatomical plausibility than about lesions. Of the four explanation methods only Grad-CAM was scored, on just 60 images and without confidence intervals, and the Grad-CAM layer we picked for EfficientNetB4 was unsuitable, so its XAI scores cannot be compared with the others. The OOD test used only synthetic Gaussian and uniform noise, and the fairness audit used simulated demographic labels instead of verified records; both are demonstrations of method, not final answers. With no patient identifiers in the source dataset, we could not check that splits are independent at the patient level, nor rule out near duplicate images. Model comparisons used an unpaired test because predictions for each image were not stored. We could not reliably measure the gain from TTA over a single view, which is why the ablation starts from Xception with TTA, and the augmentation ablation and TTA sensitivity check mentioned in Methods were not carried out. Lastly, all results come from one public benchmark, and only data from other institutions can show how far they carry over.

Future work

Three pieces of work would take this further. With real demographics and images from more than one hospital, the fairness audit could rest on actual patient diversity instead of epidemiological priors. Tumor masks drawn by radiologists would make the localization and faithfulness scores clinically meaningful. And a prospective reader study, with radiologists using the XAI outputs on real cases, would show whether the framework reduces diagnostic uncertainty in practice, which is the test that matters for deployment.

Methods

Input to the pipeline is a single brain MRI slice; output is one of four labels. Six backbones pretrained on ImageNet (five CNNs and a Vision Transformer) are each trained separately under five-fold cross-validation, and an Out-of-Fold meta-learner merges their predictions. Everything else in the pipeline exists to judge the ensemble on more than accuracy: several explainability methods, calibration, out-of-distribution detection, Monte Carlo uncertainty and a demographic fairness evaluation.

Figure 1 shows the full workflow.

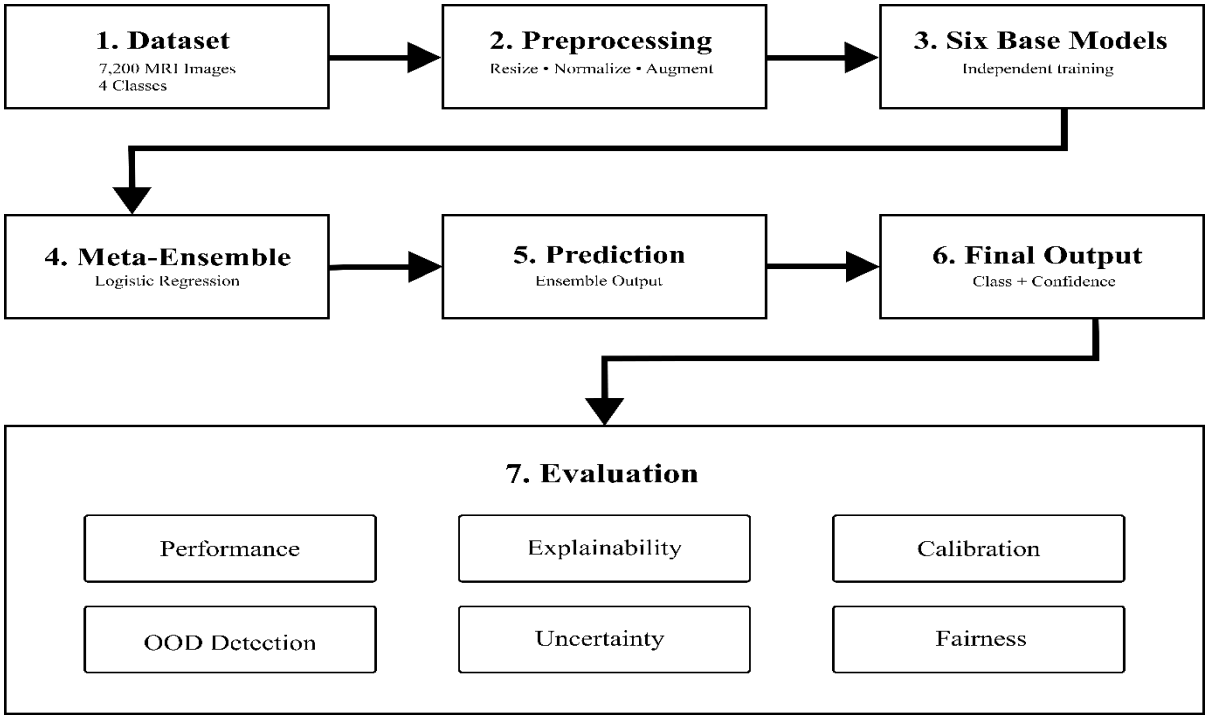


Figure 1. Overview of the proposed framework.

Figure 1 shows seven stages. Images (7,200, four classes) are resized, normalized and augmented, and six base models are trained independently. Their outputs go to a logistic regression meta-learner, and the MetaEnsemble returns a class with its confidence. In the last stage the system is evaluated along six lines: performance, explainability, calibration, OOD detection, uncertainty and fairness. Data splits, cross-validation and Out-of-Fold stacking are explained in the subsections that follow.

Dataset

We used the Brain Tumor MRI Dataset that Nickparvar released on Kaggle[37]. It contains brain MRI slices labelled glioma, meningioma, no tumor or pituitary tumor, and, as Figure 6 shows, the slices come from more than one sequence and more than one imaging plane. In its original form the release holds 7,023 images in classes of unequal size. For this study it was balanced to 1,800 images per class, 7,200 in all. Because every class has the same 1,800 images, we left the loss unweighted. All six architectures were trained with a WeightedRandomSampler using class-balanced weights. On a balanced training split these weights are identical, and the sampler simply draws images uniformly with replacement; it neither shifts the expected class mix nor forces every mini-batch to be exactly balanced.

Table 3. Dataset composition and stratified partition.

Tumor Class	Train (70%)	Val (15%)	Test (15%)	Total Images
Glioma	1,260	270	270	1,800
Meningioma	1,260	270	270	1,800
No Tumor	1,260	270	270	1,800

Tumor Class	Train (70%)	Val (15%)	Test (15%)	Total Images
Pituitary Tumor	1,260	270	270	1,800
Total	**5,040**	**1,080**	**1,080**	**7,200**

Each class is split equally across the training, validation and test sets.

Every class contributes 1,260 training, 270 validation and 270 test images. With 270 test images per class, overall accuracy and macro-averaged recall are the same number by construction, which is why they coincide in Table 5.

The training and testing folders of the Kaggle release were merged and balanced to 1,800 images per class. We then made a single stratified random split of the 7,200 images before training anything: 70% for training (n = 5,040), 15% for validation (n = 1,080) and 15% for testing (n = 1,080), or 1,260, 270 and 270 images per class. The validation split was put aside at that point and was not used in any of the experiments reported here; model selection was done through the five-fold cross-validation described below. Because the Kaggle release has no patient or scan identifiers, we cannot confirm that a patient's images fall into only one split, and exact or near duplicates across splits cannot be excluded. The Discussion returns to this. Strictly, duplicates and very similar images should be removed before splitting, and only then can the test results be called free of leakage. The test set itself was kept aside and used once, after training and meta-learner fitting had finished. We held to this strictly, because even a casual look at test images during development tends to flatter the numbers, and the model then underperforms once deployed.

Image preprocessing

Images were converted to three-channel RGB and resized for each architecture: 299 × 299 pixels for Xception and InceptionV3, 380 × 380 for EfficientNetB4, and 224 × 224 for DenseNet121, ConvNeXtTiny and ViT-B/16. We normalized pixels with the ImageNet channel statistics (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]). All six backbones were pretrained on ImageNet, so matching its statistics spares their first layers a shift in input distribution.

Training images were augmented afresh in every epoch. First came a random crop from a copy padded by 36 pixels, then horizontal flips ($p = 0.5$), vertical flips ($p = 0.2$) and rotations of up to ±25°. RandAugment followed, with two operations at magnitude 9, along with color jitter (brightness 0.35, contrast 0.35, saturation 0.25, hue 0.08). Last came a random affine transform (shear 12°, scale range [0.85, 1.15]), random perspective distortion ($p = 0.30$) and random erasing ($p = 0.20$). Some of these, namely vertical flips, perspective distortion and the wider affine and color ranges, produce images that no scanner would. We used them purely as regularization that leaves the labels unchanged. They were switched off at validation, test and TTA time (see Test-time augmentation), and we do not expect the model to cope with flipped or heavily recolored scans in practice. We have not yet run an ablation of the individual transforms to confirm that none of them hurts feature realism; that remains to be done. MixUp ($\alpha = 0.4$) was applied to 40% of training batches and CutMix ($\alpha = 1.0$) to another 35%, leaving 25% untouched. Validation and test images were only resized and normalized.

Model selection and architecture

We picked the six architectures because they see images differently. What matters for stacking is diversity rather than the number of models: networks with similar designs make similar mistakes, and combining them gains little. Four design families are represented: inception-style multiscale

convolution in Xception and InceptionV3, compound scaling in EfficientNetB4, dense skip connections in DenseNet121 and global self-attention in ViT-B/16. ConvNeXtTiny falls between the last two, being a convolutional network that borrows its normalization and several other ideas from transformers.

Architecture details

Xception[31] (21.9 M parameters, 299 × 299 input) stacks 36 convolutional layers, all depthwise separable, so it needs far fewer weights than standard convolutions of the same width. We took its last separable block (act4) as the Grad-CAM target.

InceptionV3[32] (23.0 M parameters, 299 × 299 input) replaces each n × n convolution in its inception modules with a 1 × n and n × 1 pair. That saves computation and lets a single module look at the image at several scales. We read Grad-CAM from Mixed_7c, its last inception module before global pooling.

EfficientNetB4[33] (18.6 M parameters, 380 × 380 input) comes from compound scaling: one coefficient enlarges depth, width and input resolution in fixed proportion, which buys more accuracy per FLOP than growing any one of them. Its 380-pixel inputs produced larger gradient norms, so we fine-tuned it at 3×10^{-5}, a lower rate than the others.

DenseNet121[34] (7.6 M parameters, 224 × 224 input) connects each layer inside a dense block to every layer after it. Features are reused directly and vanishing gradients become much less of a problem. Although it is the smallest model in our ensemble, its dense connectivity gives it a different inductive bias, and that difference is useful to the meta-learner.

ConvNeXtTiny[35] (28.3 M parameters, 224 × 224 input) modernizes the CNN with design choices taken from transformers: inverted bottleneck blocks, depthwise convolutions with 7 × 7 kernels, layer normalization in place of batch normalization, and GELU activations. It competes with vision transformers yet keeps the translation equivariance of convolution. Its final stage block (stages.3.blocks.2) was the Grad-CAM target.

ViT-B/16[36] (86.6 M parameters, 224 × 224 input) cuts each image into 14 × 14 = 196 separate patches of 16 × 16 pixels and projects every patch to a 768-dimensional embedding. With a learnable class token added, the 197 tokens pass through 12 transformer encoder blocks, each with 12 attention heads. Self-attention lets every position interact with every other one in a single step, so long range dependencies are captured that a local receptive field would miss. We fine-tuned it with layer-wise learning rate decay ($\gamma = 0.85$), giving blocks near the input smaller learning rates so that the pretrained spatial features were not forgotten. Grad-CAM needs feature maps of shape (B, C, H, W). We therefore dropped the class token from the output of blocks.11.norm1 and reshaped that output from (B, 197, 768) to (B, 768, 14, 14); otherwise Grad-CAM would produce a meaningless one-dimensional signal.

Table 17. Configuration of the six architectures.

Model	timm Name	Input	Batch	LR	Grad-CAM Layer	Params
Xception	xception	299	24	5×10^{-5}	act4	21.9 M
InceptionV3	inception_v3	299	24	5×10^{-5}	Mixed_7c	23.0 M
EfficientNetB4	efficientnet_b4	380	16	3×10^{-5}	conv_head	18.6 M
DenseNet121	densenet121	224	48	5×10^{-5}	features.norm5	7.6 M
ConvNeXtTiny	convnext_tiny	224	48	5×10^{-5}	stages.3.blocks.2	28.3 M
ViT-B/16	vit_base_patch16_224	224	24	2×10^{-5} †	blocks.11.norm1	86.6 M

† ViT-B/16 uses layer-wise learning rate decay (γ = 0.85). The rate shown is the base value for the final transformer block.

We set input size, batch size and learning rate per architecture according to memory. EfficientNetB4, with the largest input (380 × 380), had to use the smallest batch (16), while the 224-pixel convolutional models could take batches of 48. For every model the Grad-CAM target is the last layer that still keeps spatial structure, since a saliency map is meaningless without it.

Shared classification head

All six backbones lose their original classifier and receive the same four-class head, which Figure 11 draws layer by layer. In order, it is Global Average Pooling → Batch Normalization → Dropout (p = 0.35) → Linear (d → 512) → GELU → Batch Normalization → Dropout (p = 0.35) → Linear (512 → 256) → GELU → Dropout (p = 0.175) → Linear (256 → 4), with a softmax at the end. The two dropout layers at p = 0.35[45] earn their place twice: in training they drop random units, and at test time we leave them switched on to draw the Monte Carlo samples described under Uncertainty quantification. GELU replaced ReLU because its gradients are smoother and it has worked well in transformers and recent CNNs. With one shared head, any gap between models has to come from the backbone.

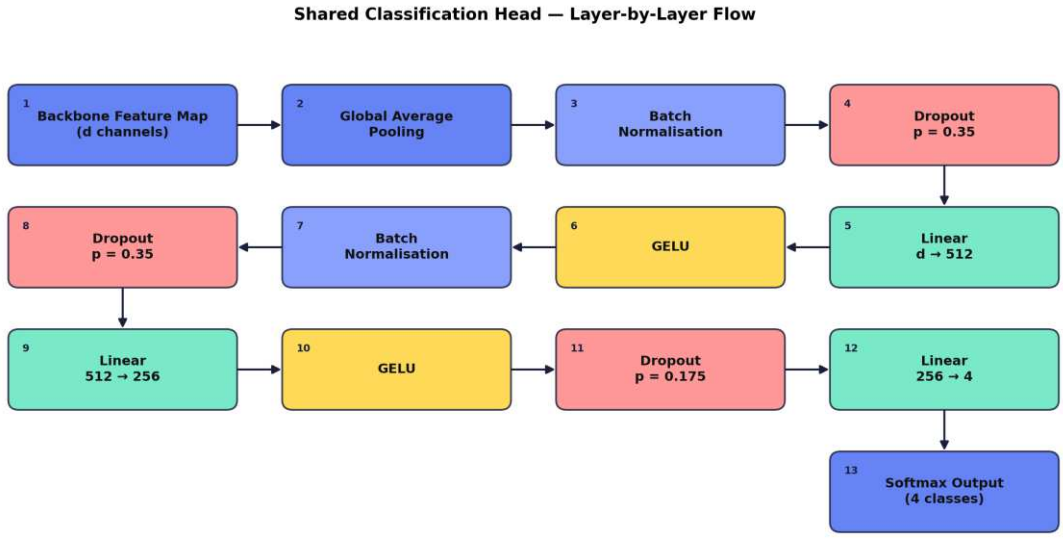


Figure 11. Layer-by-layer flow of the shared classification head applied to all six backbones.

Model training

Two-phase training protocol

Training ran in two phases. In Phase 1 (head warm-up, 5 epochs) the backbone was frozen and only the new head learned. The head thus started Phase 2 with reasonable weights, and the pretrained features were spared the large gradients of the first updates. The optimizer was AdamW, with an initial learning rate of 4×10^{-4} and weight decay of 1×10^{-4}. A linear warm-up took the learning rate from 5% of its target to the full value over three epochs, after which cosine annealing with warm restarts took over.

Phase 2 unfroze everything and trained backbone and head together for at most 55 epochs. Training stopped early if the held-out fold's validation AUC-ROC failed to improve for 12 epochs, a patience of roughly a fifth of the epoch budget. We trained in bfloat16 automatic mixed precision, which on Ampere GPUs roughly halves memory use and speeds things up without a meaningful loss of precision. Every step clipped the gradient norm at 1.0, mainly to protect the first fine-tuning epochs from gradient spikes.

Loss function and regularization

Every model was trained with Focal Loss[50] ($\gamma = 2.0$) combined with label smoothing ($\varepsilon = 0.08$). The focal term scales each example's cross-entropy by $(1 - p_c)^{\gamma}$, so an image already predicted at $p_c = 0.9$ keeps only 1% of its loss while one at $p_c = 0.5$ keeps 25%, and the gradient ends up dominated by hard and misclassified images. With c the true class and p_c its predicted probability, the loss is:

$$FL(p_c) = -(1 - p_c)^{\gamma} \cdot \log(p_c) \qquad (1)$$

Label smoothing turns the one-hot target into a soft one, giving the true class $(1 - \varepsilon)$ and each of the other classes $\varepsilon / (K - 1)$, where $K = 4$. With $\varepsilon = 0.08$ that means 0.92 for the correct class and about 0.027 for each of the other three, which stops the network from pushing its outputs all the way to 0 and 1 on training images and usually helps calibration later.

Exponential moving average

We also kept an exponential moving average (EMA) of the model parameters, with decay $\tau = 0.9998$. After each step t, the EMA parameters $\bar{\theta}$ are updated by Polyak averaging:

$$\bar{\theta}_t = \tau \cdot \bar{\theta}_{(t-1)} + (1 - \tau) \cdot \theta_t \qquad (2)$$

where θ_t denotes the current model parameters. With $\tau = 0.9998$, the average effectively spans the last 5,000 or so updates. Validation metrics, Out-of-Fold probabilities and test predictions all came from these averaged weights, which smooth out the jitter of stochastic gradient descent between steps and usually generalize better than the weights from the final step.

Five-fold cross-validation and OOF meta-learner

Cross-validation was five-fold and stratified, run separately for each architecture on the 5,040 training images (Table 3), with each fold taking its turn as the held-out part while the other four trained the model. The held-out fold supplied the validation AUC-ROC for early stopping and for picking the best EMA checkpoint. As noted earlier, the validation split of 1,080 images in Table 3 was never used: not for gradient updates, not for model selection and not for fitting the meta-learner. The 1,080 test images stayed untouched until final inference. This is why the Out-of-Fold matrix below has exactly 5,040 rows. Each training image is predicted once, by a model that did not see it during training.

We stacked the Out-of-Fold predictions of the six models column by column to form a meta-feature matrix X_{oof} of size $5{,}040 \times 24$ (six models × four class probabilities). A logistic regression meta-learner fitted to X_{oof} then learned how best to combine the models linearly, following the idea of stacked generalization[38]. Because X_{oof} holds only Out-of-Fold predictions, the meta-learner never saw test data. There is one caveat. Each held-out fold also guided early stopping and checkpoint choice, so its Out-of-Fold predictions may be slightly optimistic, although the test set was not involved. At inference the six models produced predictions on the test set averaged over the TTA views, which formed X_{test} ($1{,}080 \times 24$); the fitted meta-learner converted these into the final predictions.

Test-time augmentation

At test time each image was passed through ten fixed transformations and the softmax outputs were averaged. Our ten views were the original image; horizontal and vertical flips; rotations of 90°, 180° and 270°; a center crop with 24-pixel padding; a horizontal flip followed by a 90° rotation; and the top-left and center FiveCrop views. No patient lies in a scanner the way a vertical flip or a 90° or 270°

rotation would suggest. Those views are there only to average the output over symmetries the model has already learned; we are not claiming it should handle such orientations in a clinic. A fair check would rerun TTA with only the anatomically plausible views (original, horizontal flip, 180° rotation, center crop and center FiveCrop) and compare, and that check is not part of this submission. Averaging views ought to steady the predictions, but we have no reliable single-view baseline to say by how much (see Discussion), so all results are reported with TTA.

Evaluation metrics

Classification performance

Accuracy is reported alongside macro-averaged F1-score, precision, recall and AUC-ROC. Macro averaging scores each class on its own and then takes the unweighted mean; with four equally sized test classes this changes little here, but it keeps our numbers comparable with studies on unbalanced data. For K = 4 classes, each scored one versus rest, the macro AUC-ROC is:

$$AUC_{macro} = (1/K) \cdot \Sigma_k AUC(y_k, \hat{p}_k) \qquad (3)$$

where y_k marks whether a case belongs to class k and $\hat{p}_k$ is the probability the model gives that class. The macro F1-score is:

$$F1_{macro} = (1/K) \cdot \Sigma_k (2 \cdot Precision_k \cdot Recall_k) / (Precision_k + Recall_k) \qquad (4)$$

Calibration metrics

Calibration was measured with the Expected Calibration Error (ECE) and the Brier score. ECE splits the predictions into M = 15 confidence bins of equal width (about 0.067 each), measures in each bin how far mean confidence is from accuracy, and averages these gaps weighted by bin size[47]:

$$ECE = \Sigma_{(m=1)}^{M} (|B_m| / n) \cdot |acc(B_m) - conf(B_m)| \qquad (5)$$

where |B_m| counts the samples in bin m, n is the total number of samples, acc(B_m) is the share of the bin classified correctly and conf(B_m) is the bin's average confidence. ECE = 0 means confidence and accuracy agree in every bin. The multiclass Brier score instead squares the difference between the predicted probability vector and the one-hot label, averaged over samples:

$$BS = (1/n) \cdot \Sigma_i \Sigma_k (\hat{p}_{(i,k)} - y_{(i,k)})^2 \qquad (6)$$

Out-of-distribution detection metrics

We judge out-of-distribution (OOD) detection with three numbers. AUROC is the ROC area for telling in-distribution samples from OOD ones (1.0 is perfect separation), and AUPR is the same idea computed on the precision and recall curve. FPR95 asks how many in-distribution samples are wrongly flagged once the threshold is set to catch 95% of the OOD samples, so a low value is what we want. Two scores were compared. Maximum Softmax Probability (MSP)[48] treats the highest class probability as confidence, and the Energy score[49] is defined as:

$$E(x) = -T \cdot \log \Sigma_k \exp(f_k(x) / T) \qquad (7)$$

where f_k(x) is the k-th logit and T is a temperature parameter set to 1.0. In-distribution images usually receive lower, more negative energies. For a first sanity check we generated 200 synthetic OOD images, 100 of Gaussian noise and 100 of uniform random pixels, both normalized with ImageNet statistics. Noise of this kind is easy to spot. It is not what a deployed model would actually face, which is more

likely to be a different scanner, protocol or hospital, an imaging artifact, or a pathology outside the four classes. Testing those shifts requires external data that we did not have. Until such tests are done (see Discussion), the OOD results here should not be taken as evidence of clinical robustness.

Quantitative XAI metrics

We scored the Grad-CAM saliency maps with three metrics on a fixed set of 60 test images per architecture, sampled with stratification as 15 images per class from the held-out test split. All architectures were scored on the same 60 images, so the values in Table 12 are comparable across models. With only 60 images, though, sampling uncertainty is real, and we did not compute confidence intervals or paired comparisons for this set. The differences in Table 12 should therefore be read as indications rather than firm rankings. Localization Accuracy measures how much of the saliency energy lies inside the brain foreground, which we obtain by Otsu thresholding; saliency energy here means pixels whose activation is above 50% of the map maximum. What the metric really checks is anatomical plausibility, since the reference region is the whole brain and not a lesion mask. A map that lit up healthy tissue next to the tumor would still score well. We keep the name so that it lines up with the Pointing Game and AOPC metrics reported next to it, but it does not show localization at the lesion level, which would require tumor masks drawn by radiologists (see Discussion). The Pointing Game[43] is simpler: only the single hottest pixel counts, scoring 1 if it falls inside the foreground and 0 if it does not. AOPC faithfulness, the Area Over the Perturbation Curve[44], masks the top 10%, 20%, 30%, 40% and 50% most salient pixels one step at a time and averages the fall in predicted class confidence over the five steps:

$$AOPC = (1/5) \cdot \Sigma_{(k \in \{10\%,\ldots,50\%\})} [\, f_c(x) - f_c(x_{mask}^k) \,] \qquad (8)$$

where f_c(x) is the model's confidence in the predicted class for the original image and f_c(x_mask^k) is the confidence after the top k% of salient pixels are masked. A high AOPC means the highlighted regions actually drive the prediction rather than merely accompanying it.

Statistical significance testing

We used three statistical procedures to judge whether differences could be due to chance. The first is a 95% Wilson score interval for each model's test accuracy, which suits accuracy on a fixed test set (n = 1,080) because it is a binomial proportion. The second compares the MetaEnsemble with each single architecture through a two-proportion z-test with pooled standard error. Both accuracies are measured on the same 1,080 images, and an unpaired test ignores the fact that two models tend to fail on the same images; the test is therefore conservative. McNemar's test would be more powerful, but it needs predictions for each image, and those were not stored. We consider it the main analysis to add in future work rather than an optional extra. The third is for the fairness results. Since the fairness evaluation is repeated for six architectures, we used a nonparametric sign test to ask whether the direction of each gap, that is, which sex or age group scores lower, agrees across models more often than chance would allow. Agreement of this kind can hint at a systematic effect even when no single model has enough subgroup data for significance. The limits of this test are discussed with the fairness results.

Explainability methods

Gradient-weighted class activation mapping (Grad-CAM)

Grad-CAM[39] produces a coarse heatmap of the regions that most influenced the score for class c. The gradient of that score is taken with respect to the activations A^k of the target layer and then averaged over space, which gives the importance weights α_k^c:

$$\alpha_k^c = (1 / Z) \cdot \Sigma_i \Sigma_j \partial y^c / \partial A^k_{(i,j)} \quad (9)$$

The heatmap is a weighted sum of the feature maps, passed through a ReLU:

$$L^c_Grad\text{-}CAM = ReLU(\Sigma_k \alpha_k^c \cdot A^k) \quad (10)$$

The ReLU keeps only features that increase the class score. We upsample the map bilinearly to the input size and overlay it on the image. For ViT-B/16 the output of blocks.11.norm1 was first reshaped from (B, 197, 768) to (B, 768, 14, 14).

Grad-CAM++

Grad-CAM++[40] changes how the importance weights are computed. It uses a weighted average of positive second-order partial gradients rather than first-order ones, and its maps are sharper and more complete, especially when an object appears more than once or fills only a small part of the image. Grad-CAM++ ran on the same target layer as Grad-CAM but as a separate pass, which lets us lay the two maps over each other image by image.

$$\alpha_k^{c,ij} = (\partial^2 y^c/\partial(A^k_{ij})^2) / (2\cdot\partial^2 y^c/\partial(A^k_{ij})^2 + \Sigma_a \Sigma_b A^k_{ab} \cdot \partial^3 y^c/\partial(A^k_{ij})^3) \quad (11)$$

where α_k^{c,ij} is the second-order weight for activation unit (i, j) of feature map A^k. Multiplying each weight by the ReLU of the first-order gradient and summing over (i, j) gives the channel weight:

$$w_k^c = \Sigma_i \Sigma_j \alpha_k^{c,ij} \cdot ReLU(\partial y^c/\partial A^k_{ij}) \quad (12)$$

The Grad-CAM++ map is then formed as in Grad-CAM, but with these second-order weights replacing α_k^c:

$$L^c_{Grad\text{-}CAM++} = ReLU(\Sigma_k w_k^c \cdot A^k) \quad (13)$$

SHAP (SHapley Additive exPlanations)

SHAP[41] attributes the output to input features using Shapley values from cooperative game theory, which gives a principled account of each feature's contribution. We used GradientExplainer, which combines backpropagation with SHAP so that pixel attributions can be estimated efficiently. Sixteen background images per architecture provided the expected prediction when features are absent, and attributions were averaged over the RGB channels into one two-dimensional map per class.

For a model f, input x and feature set F, the Shapley value of feature i is:

$$\varphi_i = \Sigma_{S\subseteq F\setminus\{i\}} [|S|!\cdot(|F|-|S|-1)! / |F|!] \cdot [f_x(S\cup\{i\}) - f_x(S)] \quad (14)$$

where S is a subset of features that excludes i, and f_x(S) is the expected output when only the features in S are known and the rest are marginalized out. For images, Eq. (14) cannot be computed exactly because the number of subsets is far too large. GradientExplainer approximates the Shapley values with expected gradients instead:

$$\varphi_i \approx (x_i - x_i') \cdot E_{\alpha\sim U(0,1),\, x'\sim D} [\partial f(x'+\alpha(x-x')) / \partial x_i] \quad (15)$$

Here x′ is a background sample drawn from the reference distribution D, and α moves along the straight line from that background to the input. We estimated D from sixteen background images for each architecture and, as before, averaged φ_i across the three RGB channels.

LIME (local interpretable model-agnostic explanations)

LIME[42] explains one prediction at a time by fitting a simple linear model around it. It first cuts the image into superpixels, then makes 300 copies with random subsets of those superpixels blanked out and records what the model predicts for each copy. A weighted linear regression on these 300 pairs ranks the superpixels by influence, and we overlaid the top six on the original image.

Formally, the local explanation $\xi(x)$ is the solution of:

$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (16)$$

where f is the original model and g an interpretable surrogate from a class of linear models G. π_x weights each perturbed sample by how close it stays to x, and $\Omega(g)$ penalizes complexity so that g remains small enough to read. The loss being minimized is the locality-weighted squared error:

$$L(f, g, \pi_x) = \Sigma_{z,z' \in Z} \pi_x(z) \cdot (f(z) - g(z'))^2 \quad (17)$$

where Z is the set of 300 perturbed samples drawn for each image, z is one of them in pixel space and z' is its binary code over superpixels (present or absent).

Two different image sets were used for explainability, and the distinction matters when reading the explainability results. Qualitative panels were produced for an illustrative subset of 30 test images per architecture. Each panel shows the original MRI next to the LIME explanation, the SHAP map, the Grad-CAM and Grad-CAM++ heatmaps and an overlay averaging the normalized Grad-CAM and SHAP maps. The quantitative metrics (see Quantitative XAI metrics) used the larger set of 60 images. One set is for looking; the other is for measuring.

Uncertainty quantification

Uncertainty comes from Monte Carlo Dropout[46]. We keep the dropout layers of the shared head switched on at test time, so every forward pass samples a slightly different network, an approximate draw from the posterior. Each test image went through forty such passes for every architecture, 43,200 passes per model over the 1,080 test images. Averaging the forty softmax vectors gives the MC-averaged confidence, and the entropy of that average gives the total predictive uncertainty. In use, a case with low mean confidence or high entropy would be passed to a radiologist instead of being reported automatically.

Ethics declarations

This study used only a publicly available, de-identified dataset. No new data were collected from human participants, and ethics approval was not required.

Data availability

The Brain Tumor MRI Dataset analyzed in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (doi: 10.34740/KAGGLE/DSV/2645886). The lists of images in the balanced dataset and their split assignments are available from the corresponding author on reasonable request.

Code availability

The code is available from the corresponding author on reasonable request.

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).

2. Price, M. et al. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2018–2022. *Neuro-Oncol.* **27**, iv1–iv66 (2025).

3. Louis, D. N. et al. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncol.* **23**, 1231–1251 (2021).

4. Villanueva-Meyer, J. E., Mabray, M. C. & Cha, S. Current clinical brain tumor imaging. *Neurosurgery* **81**, 397–415 (2017).

5. Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R. & Bridge, C. P. A review of deep learning for brain tumor analysis in MRI. *npj Precis. Oncol.* **9**, 2 (2025).

6. Solanki, S., Singh, U. P., Chouhan, S. S. & Jain, S. Brain tumor detection and classification using intelligence techniques: An overview. *IEEE Access* **11**, 12870–12886 (2023).

7. Abraham, L. A., Palanisamy, G., Veerapu, G. & Nisha, J. S. Exploring the potential of explainable AI in brain tumor detection and classification: A systematic review. *Artif. Intell. Rev.* **58** (2025).

8. Hamza, A. & Damasevicius, R. Deep learning for brain tumor segmentation and classification: A systematic review of methods and trends. *Comput. Mater. Continua* **86**, 1–41 (2026).

9. Hossain, S., Chakrabarty, A., Gadekallu, T. R., Alazab, M. & Piran, M. J. Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification. *IEEE J. Biomed. Health Inform.* **28**, 1261–1272 (2024).

10. Shah, H. A. et al. A robust approach for brain tumor detection in magnetic resonance images using finetuned EfficientNet. *IEEE Access* **10**, 65426–65438 (2022).

11. Vimala, B. B., Srinivasan, S., Mathivanan, S. K., Jayagopal, P. & Dalu, G. T. Detection and classification of brain tumor using hybrid deep learning models. *Sci. Rep.* **13**, 23029 (2023).

12. Mathivanan, S. K. et al. Employing deep learning and transfer learning for accurate brain tumor detection. *Sci. Rep.* **14**, 7232 (2024).

13. Ferdous, G. J., Sathi, K. A., Hossain, M. A., Hoque, M. M. & Dewan, M. A. A. LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access* **11**, 20337–20350 (2023).

14. Afzal, S. & Lee, J. H. Efficient uncertainty-aware dual-attention network for brain tumor detection. *Mathematics* **14**, 1421 (2026).

15. Singh, R. et al. Advanced dynamic ensemble framework with explainability-driven insights for precision brain tumor classification across datasets. *Sci. Rep.* **15**, 29090 (2025).

16. Napa, K. K., Murugan, S., Senthil Murugan, J. & Jayanthi, A. BRAIN-META: A reproducible CNN–vision transformer meta-ensemble pipeline for explainable brain tumor classification. *MethodsX* **16**, 103769 (2025).

17. Nahiduzzaman, Md. et al. A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using MRI images. *Sci. Rep.* **15**, 1649 (2025).

18. Musthafa, M., Mahesh, T. R., Vinoth Kumar, V. K. & Guluwadi, S. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet-50. *BMC Med. Imag.* **24**, 190 (2024).

19. Nazir, Md. I., Akter, A., Wadud, Md. A. H. & Uddin, Md. A. Utilizing customized CNN for brain tumor prediction with explainable AI. *Heliyon* **10**, e38997 (2024).

20. Akgundogdu, A. & Celikbas, S. Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy. *Med. Eng. Phys.* **144**, 104405 (2025).

21. Ahmed, Md. M. et al. Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh. *Sci. Rep.* **14**, 22797 (2024).

22. Iftikhar, S., Anjum, N., Siddiqui, A. B., Rehman, M. U. & Ramzan, N. Explainable CNN for brain tumor detection and classification through XAI-based key features identification. *Brain Inform.* **12**, 10 (2025).

23. Mehta, R. et al. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation. *Mach. Learn. Biomed. Imag.* **1**, 1–54 (2022).

24. Kendall, A. & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) 30, 5574–5584 (2017).

25. Zou, K. et al. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiol.* **1**, 100003 (2023).

26. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) 30, 6402–6413 (2017).

27. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, 4581 (2022).

28. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).

29. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: A modelling study. *Lancet Digit. Health* **4**, e406–e414 (2022).

30. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Nat. Acad. Sci. USA* **117**, 12592–12594 (2020).

31. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 1251–1258 (2017).

32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception architecture for computer vision. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2818–2826 (2016).

33. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proc. 36th Int. Conf. Mach. Learn. (ICML) 6105–6114 (2019).

34. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 4700–4708 (2017).

35. Liu, Z. et al. A ConvNet for the 2020s. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) 11976–11986 (2022).

36. Dosovitskiy, A. et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proc. Int. Conf. Learn. Represent. (ICLR) (2021).

37. Nickparvar, M. Brain tumor MRI dataset. Kaggle https://doi.org/10.34740/KAGGLE/DSV/2645886 (2021).

38. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).

39. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

40. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV) 839–847 (2018).

41. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) 30, 4765–4774 (2017).

42. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?': Explaining the predictions of any classifier. In Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD) 1135–1144 (2016).

43. Zhang, J. et al. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.* **126**, 1084–1102 (2018).

44. Samek, W., Binder, A., Montavon, G., Lapuschkin, S. & Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 2660–2673 (2017).

45. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

46. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proc. 33rd Int. Conf. Mach. Learn. (ICML) 1050–1059 (2016).

47. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In Proc. 34th Int. Conf. Mach. Learn. (ICML) 1321–1330 (2017).

48. Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proc. Int. Conf. Learn. Represent. (ICLR) (2017).

49. Liu, W., Wang, X., Owens, J. D. & Li, Y. Energy-based out-of-distribution detection. In Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) 33, 21464–21475 (2020).

50. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In Proc. IEEE Int. Conf. Comput. Vis. (ICCV) 2980–2988 (2017).

Author contributions

Md. Sajjad Ullah: Methodology, Coding. Shochi Akter: Drafting & Editing. Sadman Sadik Khan: Supervision. All authors read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Use of generative AI

During the preparation of this manuscript, the authors used Claude (Anthropic) to improve the language and readability of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Additional information

Correspondence and requests for materials should be addressed to S.S.K.