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.

Atanassov Intuitionistic Fuzzy Classification of Brain Tumor Patches in BraTS FLAIR MRI: Accuracy, Hesitation and Explainability

Rogi Jacob

rogijacob@gmail.com

Belgium Campus It versity

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11766897/v1

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

Additional Declarations: No competing interests reported.

Atanassov Intuitionistic Fuzzy Classification of Brain Tumor Patches in BraTS FLAIR MRI: Accuracy, Hesitation and Explainability

Rogi Jacob[1*]

[1*]Department of Mathematics, Belgium Campus, Pretoria, South Africa.

Corresponding author(s). E-mail(s): Jacob.r@belgiumcampus.ac.za;

Abstract

The proposed work uses a feature-based Atanassov intuitionistic fuzzy classifier (AIFC) for patch-level whole-tumor detection in FLAIR MRI from BraTS 2020. Each $\mathbf{11 \times 11}$ patch is described by four intensity moments and a center-referenced maximum deviation, and a trainable layer converts them into membership, non-membership and hesitation degrees before a small multilayer perceptron. Across five patient-level splits of 369 patients, the classifier reached an AUC of $\mathbf{0.979 \pm 0.005}$ and a Dice score of $\mathbf{0.83 \pm 0.02}$ on unseen patients, indistinguishable from a plain MLP on the same features. SHAP, LIME and permutation importance consistently show that mean intensity accounts for roughly 83% to 97% of the model's importance, and a logistic regression on mean intensity alone comes within about 0.005 AUC and 0.014 F1. Hesitation is lowest inside tumor tissue because it follows mean intensity, and it neither marks the tumor boundary nor flags errors. On this task the intuitionistic layer therefore adds readable parameters but no gain in accuracy or uncertainty, and simple baselines and attribution checks should accompany fuzzy neural classifiers.

Keywords: Atanassov intuitionistic fuzzy sets, brain tumor, BraTS, MRI, explainable artificial intelligence, SHAP, LIME

1 Introduction

Deep learning has become the standard approach to analyzing brain MRI. Convolutional networks such as U-Net, ResNet and nnU-Net perform well in tumor detection,

classification and segmentation [1, 3–6, 16, 18, 20, 21], and more recently attention-based models have been added to the toolbox [25]. Public benchmarks, above all the BraTS challenge, have driven this progress [17, 23, 26], and in some medical imaging tasks the accuracy of such systems has matched that of experts [7]. Most of these models are trained end to end on large annotated image sets, often with data augmentation [2, 8], and they are hard to inspect: a clinician receives a label but little information about which image properties produced it or how uncertain the decision is [22]. Two properties of MRI make uncertainty hard to ignore. Images from different scanners and centers differ in contrast, resolution and acquisition protocol [9, 10], and tissue is heterogeneous, so the boundary between tumor, edema and healthy tissue is often gradual. Fuzzy-set methods were introduced to describe such imprecision, and neuro-fuzzy systems combine them with learning [13–15]. Atanassov intuitionistic fuzzy sets [11, 12] extend a membership degree with a non-membership degree and a hesitation degree, which could in principle express how much evidence for and against a class is missing. Intuitionistic fuzzy methods have been applied to brain MRI segmentation and, more recently, inside deep networks [27, 28]. The claim that the hesitation degree gives useful, interpretable uncertainty is, however, rarely tested directly: studies seldom compare against simple baselines, split the data by patient, or examine what the trained model actually uses. This paper builds a compact intuitionistic fuzzy classifier and tests those claims. Each 11×11 FLAIR patch is described by four intensity moments and a local contrast measure. A trainable layer converts these features into membership, non-membership and hesitation degrees, and a small neural network classifies the result. We evaluate the classifier on 369 BraTS 2020 patients with patient-level splits, compare it with a plain neural network and with simple baselines, and explain the models with SHAP [22], LIME [24] and permutation importance. The contributions are:

1. A trainable intuitionistic fuzzy layer with feature-specific memberships, a Sugeno-type non-membership whose parameter depends on local contrast, and learnable aggregation weights.
2. A patient-level evaluation over five random splits, which shows that the classifier matches a plain MLP (AUC 0.979, Dice 0.83) and that a one-feature model comes within about 0.005 AUC and 0.014 F1.
3. An explainability analysis showing that SHAP, LIME and permutation importance agree that mean FLAIR intensity accounts for most of the model's evidence.
4. An analysis of the hesitation output, which tracks tumor class through mean intensity and does not mark the tumor boundary and detects errors worse than the network's own confidence. We report this as a negative finding.

2 Background: intuitionistic fuzzy sets

An Atanassov intuitionistic fuzzy set A on a universe X assigns to each element x a membership degree $\mu_A(x)$ and a non-membership degree $\nu_A(x)$ in $[0, 1]$ with $\mu_A(x) +$

$\nu_A(x) \leq 1$ [11, 12]. The hesitation degree

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \in [0, 1] \tag{1}$$

measures the evidence that is neither for nor against membership. If $\pi_A(x) = 0$ for every x, A is an ordinary fuzzy set. An intuitionistic fuzzy value can be generated from a fuzzy membership μ with a fuzzy complement. The Sugeno complement $\nu = (1 - \mu)/(1 + \lambda\mu)$, with $\lambda > 0$, satisfies $\nu \leq 1 - \mu$ and gives $\pi = \lambda\mu(1 - \mu)/(1 + \lambda\mu)$. This hesitation is zero at $\mu = 0$ and $\mu = 1$ and largest at intermediate membership, and it is the form used here. Several intuitionistic fuzzy values can be combined with the intuitionistic fuzzy weighted average, which preserves the constraint $\mu + \nu \leq 1$. The patch features are the first four statistical moments of the gray-level distribution (mean, variance, skewness, excess kurtosis) and a local contrast measure that Section 3 defines as the Hausdorff distance between the patch intensity set and its center pixel.

3 Methods

3.1 Data and task

We use the BraTS 2020 training set [17, 23, 26]: 369 patients, each with a FLAIR volume and a tumor segmentation. Each non-overlapping 11×11 patch is classified as tumor (1) or non-tumor (0). A patch is positive if more than half of its pixels carry any BraTS tumor label (whole tumor). The output is a coarse patch-level map, not a pixel-accurate segmentation. After filtering, the data hold 211,835 patches, of which 17,509 (8.3%) are tumor.

3.2 Preprocessing and features

- **Normalization.** Intensities are z-scored per volume, using the non-zero (brain) voxels of the whole volume.
- **Slices.** Six evenly spaced slices between the 20th and 80th percentile of slices containing brain, chosen from the image only and never from the ground truth. Tumor-free slices are included.
- **Patch filter.** A patch is kept if at least 90% of its pixels are brain and its intensity standard deviation is at least 0.02, which removes skull-edge and flat patches.
- **Features.** For a patch with $m = 121$ pixels $x_1, \ldots, x_m$ and center pixel x_c: the mean, the variance, the skewness, the excess kurtosis, and $D = \max_k |x_k - x_c|$.
- **Scaling.** Features are standardized with training-set statistics and clipped to ± 5.

D is the Hausdorff distance between the patch intensity set and the singleton $\{x_c\}$. Because x_c belongs to the patch, the reverse directed distance is 0, so D reduces to the largest intensity gap to the center pixel. It is a local contrast measure computed on intensities, not on coordinates, and it describes no shape. D correlates strongly with the variance (Spearman 0.91).

3.3 AIFS membership layer

The layer takes the five standardized features z and returns 15 values: μ_j, ν_j and π_j for the four moments, and the aggregated μ, ν, π. Mean and variance use a sigmoid membership; skewness and kurtosis use a two-sided Gaussian:

$$\mu_j = \frac{1}{1 + e^{-a_j(z_j - c_j)}}, \qquad j \in \{\text{mean}, \text{variance}\}, \tag{2}$$

$$\mu_j = \exp\left(-\frac{(z_j - c_j)^2}{2s_j^2}\right), \quad s_j = \begin{cases} s_j^L, & z_j < c_j \\ s_j^R, & z_j \geq c_j \end{cases}, \qquad j \in \{\text{skewness}, \text{kurtosis}\}. \tag{3}$$

Non-membership comes from the Sugeno generator, with a parameter λ that depends on the local contrast $\tilde{D}$ (standardized D):

$$\nu_j = \frac{1 - \mu_j}{1 + \lambda\,\mu_j}, \qquad \pi_j = 1 - \mu_j - \nu_j, \qquad \lambda = \text{softplus}(\lambda_0 + \lambda_1 \tilde{D}) + 10^{-3}. \tag{4}$$

The four feature-level triples are combined with the intuitionistic fuzzy weighted average, which preserves $\mu + \nu \leq 1$:

$$\mu = 1 - \prod_j (1 - \mu_j)^{w_j}, \qquad \nu = \prod_j \nu_j^{w_j}, \qquad \pi = 1 - \mu - \nu, \qquad w = \text{softmax}(\omega). \tag{5}$$

Memberships are clipped to $[10^{-4}, 1 - 10^{-4}]$. The centers c, slopes a, widths s, λ_0, λ_1 and weights w are trained with the classifier. The variant `AIFS_noD` freezes λ_1 at 0, so λ no longer depends on D.

3.4 Classifier and training

The 15 layer outputs feed a network with a 32-unit ReLU layer, dropout 0.3, a 16-unit ReLU layer and one sigmoid output. Training uses Adam (learning rate 10^{-3}), binary cross-entropy, batch size 128, at most 60 epochs, and early stopping on validation AUC with patience 6 (best weights restored). Three variants are compared: `MLP_raw` (the same network on the five standardized features, with no AIFS layer), `AIFS_full`, and `AIFS_noD`.

3.5 Splits, class imbalance and thresholds

- **Patient-level splits.** 239 / 56 / 74 patients for training, validation and testing; no patient appears in two sets. The split is repeated for five random seeds.
- **Imbalance.** SMOTE [19] is applied to the training set only; the scaler is also fitted on training data only. Validation and test sets keep the real class balance.
- **Decision rule.** A patch is called tumor if the network probability is at least τ and the aggregated hesitation is at most η. τ (0.30 to 0.99 in steps of 0.01) and η (candidate quantiles of the validation hesitation) are chosen to maximize validation F1. The variant "no η cap" drops the η rule and re-chooses τ.

3.6 Evaluation

We report recall, specificity, precision, F1 (equal to the Dice score for patch labels), balanced accuracy and ROC-AUC on the test patients, as mean ± SD over the five seeds. AUC is computed once per model from the continuous output, so it does not depend on τ or η. Variants are compared with paired differences over seeds. The qualitative figure shows three test patients on the slice with the most tumor patches, chosen for display only.

3.7 Simple baselines

Three simple models were trained on the same patient-level splits and standardized features: logistic regression on mean intensity alone, logistic regression on all five features, and histogram gradient boosting on all five features. They use class-balanced weights instead of SMOTE, and τ is chosen on the validation patients by F1, as for the networks. We also report the AUC of the raw mean intensity used directly as a score.

3.8 Explainability analysis

For three seeds, the MLP and the AIFS model were retrained on the training patients and explained on the unseen test patients.

- **SHAP.** Exact interventional Shapley values [22] for the tumor probability, obtained by enumerating all feature coalitions (this is what KernelSHAP converges to). The sample is 1,000 test patches (500 per class), and the background is 100 k-means centers of real, non-SMOTE training patches. Variance and D are treated as one group because they correlate at 0.91. Importance is the mean absolute SHAP value, averaged over the two classes with equal weight.
- **LIME.** Tabular LIME [24] on continuous features for 150 test patches with 2,000 perturbations each. A patch's contribution is the local coefficient times the standardized deviation, and 30 patches were explained a second time with a different random seed to measure stability.
- **Permutation importance.** The drop in test ROC-AUC after shuffling a feature (or the variance and D group jointly) five times on the full test set.
- **Hesitation drivers.** SHAP applied to the aggregated hesitation π of the AIFS model.
- **Comparison.** Each method's importances are normalized to sum to 1. Agreement is measured by the Spearman correlation between rankings of the four groups (a coarse measure with so few groups), and per patch between SHAP and LIME contributions together with the share of matching signs.

3.9 Error-detection analysis

For the AIFS and the MLP, a patch is an error if it is misclassified under the validation-chosen threshold without any hesitation cap. Four uncertainty scores, each higher for more uncertain patches, are compared: the aggregated hesitation π, the standard deviation of the four feature-level memberships (disagreement among the features),

the distance of the output logit from the threshold, and the binary predictive entropy. On the test patients we report the AUROC for detecting errors (0.5 is chance), the share of all errors found among the 5%, 10% and 20% most uncertain patches (chance equals that percentage), a scaled area under the risk-coverage curve (0 is random ordering, 1 is perfect), and the AUROC for telling false from true positives among the patches predicted as tumor.

3.10 Implementation

The pipeline is implemented in Python 3.14.5 on Windows 11 with NumPy 2.5.3, pandas 3.0.6, SciPy 1.18.1, scikit-learn 1.9.1, imbalanced-learn 0.14.2, nibabel 5.4.2, Matplotlib 3.11.2 and TensorFlow 2.22.0rc0 (with Keras 3.16.0.dev2026101004). Shapley values were computed with our own exact implementation, and LIME with the `lime` package (version 0.2.0.1). The code is available at https://github.com/andrewkalathil/AIFC-.

4 Results

On unseen patients the AIFS classifier matches the plain MLP: F1/Dice 0.831 versus 0.832, and AUC 0.979 for both. Table 1 gives mean ± SD over five patient-level splits (74 test patients each). Because every variant uses the same splits per seed,

Table 1 Test performance (mean ± SD over 5 seeds).

Variant	Recall	Specificity	Precision
`MLP_raw`	0.809 ± 0.022	0.988 ± 0.004	0.857 ± 0.037
`AIFS_full`	0.811 ± 0.026	0.988 ± 0.003	0.853 ± 0.028
`AIFS_full`, no η cap	0.812 ± 0.027	0.987 ± 0.003	0.852 ± 0.028
`AIFS_noD`	0.811 ± 0.015	0.987 ± 0.003	0.848 ± 0.028
`AIFS_noD`, no η cap	0.814 ± 0.018	0.987 ± 0.002	0.845 ± 0.023
Variant	F1/Dice	Balanced acc.	AUC
`MLP_raw`	0.832 ± 0.024	0.899 ± 0.012	0.979 ± 0.006
`AIFS_full`	0.831 ± 0.023	0.899 ± 0.013	0.979 ± 0.005
`AIFS_full`, no η cap	0.831 ± 0.022	0.900 ± 0.014	0.979 ± 0.005
`AIFS_noD`	0.829 ± 0.019	0.899 ± 0.009	0.979 ± 0.004
`AIFS_noD`, no η cap	0.829 ± 0.019	0.900 ± 0.010	0.979 ± 0.004

paired differences are the fairer comparison. All of them are within the seed-to-seed spread. The chosen τ lay between 0.86 and 0.94, which is expected because SMOTE balances the training classes while real prevalence is 8%. An F1-optimal threshold favors precision (0.85) over recall (0.81); if sensitivity is the clinical priority, τ should instead be set on validation data to reach a target recall, and the resulting specificity and precision reported. The test sets of different seeds overlap, so the SDs describe split variability and are not confidence intervals.

Table 2 Paired differences over seeds (first minus second).

Comparison	ΔF1 (SD)	Seeds better	ΔAUC (SD)	Seeds better
`AIFS_full` − `MLP_raw`	−0.0010 (0.0019)	2/5	−0.0003 (0.0017)	3/5
`AIFS_noD` − `MLP_raw`	−0.0032 (0.0056)	1/5	−0.0005 (0.0022)	3/5
`AIFS_full` − `AIFS_noD`	+0.0021 (0.0038)	4/5	+0.0002 (0.0005)	2/5
`AIFS_full` − `AIFS_full` without η cap	+0.0001 (0.0002)	2/5	0.0000 (0.0000)	0/5

4.1 Simple baselines

A single soft intensity threshold comes close to the neural models. The baselines use the same features, patient-level splits (seeds 0 to 4) and threshold procedure as the neural models, with class-balanced weights instead of SMOTE. Logistic regression

Table 3 Baselines versus neural models (mean ± SD over 5 seeds).

Model	Recall	Precision
Mean intensity only (logistic regression)	0.792 ± 0.033	0.847 ± 0.019
Logistic regression, 5 features	0.798 ± 0.019	0.829 ± 0.028
Gradient boosting, 5 features	0.804 ± 0.021	0.864 ± 0.033
`MLP_raw`	0.809 ± 0.022	0.857 ± 0.037
`AIFS_full`	0.811 ± 0.026	0.852 ± 0.028
Model	**F1/Dice**	**AUC**
Mean intensity only (logistic regression)	0.818 ± 0.021	0.974 ± 0.005
Logistic regression, 5 features	0.813 ± 0.018	0.976 ± 0.004
Gradient boosting, 5 features	0.833 ± 0.024	0.980 ± 0.006
`MLP_raw`	0.832 ± 0.024	0.979 ± 0.006
`AIFS_full`	0.831 ± 0.023	0.979 ± 0.005

on mean intensity alone reaches an AUC of 0.974 and an F1 of 0.818. The neural models are better in all five seeds: the MLP by 0.003 to 0.009 AUC and 0.009 to 0.020 F1, and the AIFS model by 0.003 to 0.009 AUC and 0.010 to 0.016 F1. The gain is consistent but small. Gradient boosting on the five features matches both networks (AUC 0.980 and F1 0.833, against 0.979 and 0.832 for the MLP and 0.979 and 0.831 for the AIFS model), with an average AUC difference of 0.001 that lies within the seed-to-seed variation. The AIFS layer therefore gives no advantage over a standard tree ensemble on the same features.

4.2 Hesitation by region

Hesitation is lowest inside tumor tissue and about twice as high everywhere else. It is only marginally higher at the tumor rim than in the core (0.084 versus 0.078) and at the healthy rim than in the interior (0.169 versus 0.161), so it does not localize the boundary. Brain-edge patches no longer stand out after the 90% brain-fraction filter. This pattern follows from the Sugeno form: hesitation is largest at medium

Table 4 Aggregated hesitation π by region (seed 0, test patients).

Region	Patches	Mean π	Median π	Mean D
Tumor core	842	0.078	0.060	1.88
Tumor rim	2,677	0.084	0.074	2.22
Healthy rim	4,121	0.169	0.168	1.93
Healthy interior	18,458	0.161	0.158	1.55
Brain edge	16,050	0.160	0.157	1.63

membership and falls toward both extremes. Tumor patches have high μ, so π is small, while healthy patches receive intermediate μ because the learned sigmoid slopes (about 1 to 2.5) cannot push them to 0. In this model π therefore tracks the class more than any independent uncertainty.

4.3 Hesitation as an error detector

Hesitation does not find the classifier's errors; the network's own distance to its decision threshold does (Table 5, Figure 2). A random ordering would give an AUROC of 0.5, 10% of the errors in the top 10% of patches, and a gain of 0. The 10% of

Table 5 Error detection by uncertainty score (`AIFS_full`; mean $\pm$ SD over 5 seeds, test patients).

Uncertainty score	AUROC for errors	Errors in the 10% most uncertain
Distance to the threshold	0.926 ± 0.009	0.776 ± 0.045
Predictive entropy	0.720 ± 0.036	0.240 ± 0.061
Hesitation π	0.419 ± 0.045	0.049 ± 0.021
Spread of the feature memberships	0.372 ± 0.080	0.035 ± 0.011
Uncertainty score	**Risk-coverage gain**	**AUROC, false vs true positives**
Distance to the threshold	0.891 ± 0.015	0.869 ± 0.023
Predictive entropy	0.509 ± 0.115	0.869 ± 0.023
Hesitation π	-0.228 ± 0.142	0.778 ± 0.056
Spread of the feature memberships	-0.301 ± 0.362	0.405 ± 0.040

patches closest to the threshold contain 78% of all errors, and the MLP's distance to the threshold behaves the same (AUROC 0.926). Hesitation is worse than chance overall: the 10% most hesitant patches hold only 4.9% of the errors, so errors receive lower hesitation, not higher. Among the patches predicted as tumor, hesitation does separate false from true positives (AUROC 0.78), but less well than the distance to the threshold (0.87). Disagreement among the feature memberships is also worse than chance. The Spearman correlation of hesitation with mean intensity is only -0.07, but this does not mean independence, because hesitation is a non-monotone function of the mean. A possible reason is that hesitation is largest at intermediate membership, which corresponds to mid-range intensities, whereas the classifier's errors lie around

its decision threshold (about 0.9 after SMOTE), where membership is already high and hesitation small. This is a hypothesis that the present data do not test.

4.4 Learned parameters

- Feature weights w_j (five seeds): mean 0.42 to 0.69 (average 0.60), variance 0.11 to 0.21 (0.17), kurtosis 0.10 to 0.18 (0.13) and skewness 0.04 to 0.29 (0.11; 0.04 to 0.08 in four of the five seeds). These weights should not be read as feature importance: the explainability analysis puts mean intensity at 0.86 (SHAP) to 0.97 (permutation), well above its weight, because the network also receives the per-feature memberships directly.
- λ_1 is positive in four of the five seeds (0.54 to 1.05) and negative in one (-0.48), so the layer does not consistently raise hesitation with local contrast. This agrees with SHAP applied to π, where variance and D together lower hesitation slightly.
- The mean-membership center lies 0.47 to 0.65 standard deviations above the training-patch average of the feature, and the skewness center lies between -1.02 and -0.31. The right-hand width of the kurtosis membership varies from 1.5 to 3.8 across seeds, so it is poorly determined, while the left-hand width is stable (0.68 to 0.91).

4.5 Qualitative results

Figure 3 shows, for three unseen test patients, the FLAIR slice, the ground truth, the overlay of true positives, false positives and false negatives, and the hesitation map. In Figure 3, one can easily see that tumor cores are detected, errors lie at the margins and in bright healthy tissue, and the hesitation map does not highlight them.

5 Explainability: SHAP, LIME and permutation importance

Three attribution methods agree that mean FLAIR intensity drives the classifier, and the AIFS and MLP models rely on the features in the same way (three seeds, unseen test patients, method as described in Section 3). Mean intensity carries 83% to 88% of the importance by SHAP and LIME and about 97% of the AUC loss under permutation. The other three groups together carry 3% to 17%. The learned weight of the mean (0.57) understates its role, so the weights of the AIFS layer are not a reliable explanation of the model. Figure 4 shows these shares as a chart. The group-level rank correlations are coarse because only four groups are ranked; the differences concern the order of the three minor groups, not the first place of the mean. LIME is stable across reruns.

5.1 Shape of the dependence on mean intensity

The SHAP value of the mean rises steeply between about 0 and 1.5 standard deviations above the brain mean and flattens at about $+0.7$ beyond roughly $+2$, with a floor

Table 6 Share of total importance by method (mean ± SD over 3 seeds; each method normalized to sum to 1).

Feature group	SHAP, AIFS	SHAP, MLP	Permutation, AIFS
Mean	0.864 ± 0.027	0.880 ± 0.010	0.975 ± 0.003
Variance + D	0.058 ± 0.004	0.063 ± 0.009	0.014 ± 0.004
Skewness	0.028 ± 0.011	0.018 ± 0.005	0.005 ± 0.001
Kurtosis	0.051 ± 0.019	0.039 ± 0.010	0.007 ± 0.003
Feature group	Permutation, MLP	LIME, AIFS	LIME, MLP
Mean	0.974 ± 0.007	0.837 ± 0.023	0.829 ± 0.023
Variance + D	0.017 ± 0.006	0.080 ± 0.007	0.092 ± 0.016
Skewness	0.004 ± 0.001	0.019 ± 0.003	0.019 ± 0.001
Kurtosis	0.005 ± 0.000	0.063 ± 0.027	0.061 ± 0.009
Feature group	Learned weight w_j, AIFS		
Mean	0.569 ± 0.133		
Variance + D	0.163 ± 0.050		
Skewness	0.133 ± 0.134		
Kurtosis	0.135 ± 0.044		

Table 7 Agreement between methods.

Model	SHAP vs LIME per patch (rank)	Matching signs	LIME rerun (rank)
`AIFS_full`	0.72 (0.69 to 0.74)	86% (83% to 89%)	0.94 (0.92 to 0.96)
`MLP_raw`	0.76 (0.72 to 0.78)	90% (89% to 90%)	0.97 (0.97 to 0.99)
Model	Group rank, SHAP vs permutation	Group rank, SHAP vs LIME	
`AIFS_full`	0.8 to 1.0	0.8 to 1.0	
`MLP_raw`	1.0	1.0	

near -0.25 below 0. AIFS and MLP follow the same curve, so both behave as a soft threshold on mean intensity (Figure 5).

5.2 What drives hesitation

SHAP applied to π shows that mean intensity lowers hesitation by 0.053 ± 0.015 in tumor patches and raises it by 0.019 ± 0.006 in non-tumor patches. Variance and D together shift it by -0.004 ± 0.002 and -0.009 ± 0.003, and skewness and kurtosis by less than 0.005 in absolute value. Hesitation is therefore mainly a function of mean intensity, as the region analysis suggested, and D does not raise it, even though λ_1 is positive in the three explained seeds.

5.3 Individual patches

The three example patches (one true positive, one false positive, one false negative) show the pattern directly. The false positive (probability 0.94) is pushed upward by the mean (about +0.57), as the true positive (0.97) is, and its hesitation of 0.15 is close to the true positive's 0.13. The false negative (0.04) is pulled down by a low mean

(about -0.33). Errors are therefore bright healthy tissue or dim tumor, and π does not flag them. These are single examples, shown for illustration only, in the supplement with the beeswarm plots (Appendix).

5.4 What the explanations do and do not show

They show that the model, and the plain MLP as well, depends almost entirely on mean FLAIR intensity, which fits tumor and edema being bright on this sequence. They do not show that the predictions are clinically valid. Attributions depend on the background data and are shared among correlated features (variance and D, which are grouped here for that reason). Radiologist review of error cases and an external dataset would be needed for any statement about clinical validity.

6 Discussion and limitations

Mean FLAIR intensity explains most of what the classifier does, and a one-feature model comes within about 0.005 AUC and 0.014 F1, so the contribution of this work is an analysis and not a better classifier.

- **Baselines and features.** Gradient boosting on the same five features matches both networks, and mean intensity alone is close. Skewness, kurtosis and D carry roughly 1% to 14% of the importance depending on the method, so they add little on FLAIR alone. Adding T1ce and T2 might give them more to do.
- **AIFS layer.** Accuracy equals the MLP's, and the learned weights understate the role of the mean, so they cannot be presented as feature importance.
- **Hesitation.** It follows mean intensity: lowest in tumor, about twice as high elsewhere, only marginally higher at the tumor rim than in the core, and no higher for the false positive example than for the true positive. The η cap changes F1 by 0.0001. As an error detector it is worse than chance (AUROC 0.42), and so is the disagreement among the feature memberships (0.37), while the network's own distance to its threshold is the best signal (0.93). A hesitation tied to the decision boundary, for example trained against the tumor fraction of each patch, would be a different and testable design.
- **Role of D.** Letting D drive λ changed F1 by $+0.002$ (better in 4 of 5 seeds) and AUC by $+0.0002$, differences within the seed-to-seed spread, so we draw no conclusion about an effect of D.
- **Scope.** FLAIR only, whole tumor versus non-tumor, 11×11 patches, six slices per patient, one dataset. The results are not comparable to pixel-level segmentation benchmarks.
- **Statistics.** Five random splits for performance and three for the explanations, with overlapping test patients, give a spread and not a confidence interval.
- **Claims not made.** Robustness to differences between scanners or centers, suitability for low-specification hardware, and interpretability of the tumor boundary were not evaluated, so the paper makes no claims about them. Multi-site or degraded-image experiments and timing and memory measurements would be needed to test them.

7 Conclusion

A feature-based Atanassov intuitionistic fuzzy classifier detected tumor patches in FLAIR MRI with an AUC of 0.979 and a Dice score of 0.83 on unseen BraTS 2020 patients, matching a plain MLP on the same features. SHAP, LIME and permutation importance show that both models depend almost entirely on mean FLAIR intensity (83% to 97% of importance), and a logistic regression on mean intensity alone is within about 0.005 AUC and 0.014 F1, so the intuitionistic layer adds readable memberships but no gain in accuracy. Its hesitation output follows mean intensity and neither marks the tumor boundary nor flags errors. Future work should test hesitation trained against the tumor fraction of each patch, so that it is tied to the decision boundary, multi-modal inputs and institution-level hold-out evaluation before claiming uncertainty-awareness or robustness.

Declarations

Funding. The author received no financial support from any organization for the submitted work.

Competing interests. The author declares that there are no competing interests.

Data availability. The BraTS 2020 training dataset used in this study is publicly available through the BraTS challenge [17, 23, 26].

Code availability. The source code used in this study is publicly available at: https://github.com/andrewkalathil/AIFC-.

Ethics approval. This study involved secondary analysis of publicly available BraTS 2020 medical imaging data. No new data were collected directly from human participants. The dataset was made available by the BraTS organizers. The ethical approval and consent arrangements applicable to the original dataset are those established by its data providers.

Consent to participate. Not applicable. No participants were recruited directly for this study.

Consent to publish. Not applicable. This study used publicly available, anonymised BraTS 2020 MRI data, and no personally identifiable patient information is included in this manuscript.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[2] C. Shorten, T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[3] Y. LeCun, Y. Bengio, G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[5] O. Ronneberger, P. Fischer, T. Brox. U-Net: convolutional networks for biomedical image segmentation. *MICCAI*, 2015, pp. 234–241.

[6] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. *CVPR*, 2016, pp. 770–778.

[7] A. Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[8] H. H. Sultan, N. M. Salem, W. Al-Atabany. Multi-classification of brain tumor images using deep neural network. *IEEE Access*, 7:69215–69225, 2019.

[9] R. T. Shinohara et al., "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014. doi: 10.1016/j.nicl.2014.08.008.

[10] M. Ghafoorian et al., "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*, Part III, Lecture Notes in Computer Science, vol. 10435. Cham, Switzerland: Springer International Publishing, 2017, pp. 516–524.

[11] K. T. Atanassov. Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 20(1):87–96, 1986.

[12] K. T. Atanassov. *Intuitionistic Fuzzy Sets: Theory and Applications*. Physica-Verlag, Springer, Heidelberg, 1999.

[13] G. Andonovski et al. Advancements in data-driven evolving fuzzy and neurofuzzy control: a comprehensive survey. *Applied Soft Computing*, 186, 2026.

[14] P. V. de Campos Souza, E. Lughofer. EFNN-NullUni: an evolving fuzzy neural network based on null-uninorm. *Fuzzy Sets and Systems*, 449:1–31, 2022.

[15] R. Yin et al. A rule-based deep fuzzy system with nonlinear fuzzy feature transform for data classification. *Information Sciences*, 633:431–452, 2023.

[16] Z. Akkus et al. Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of Digital Imaging*, 30(4):449–459, 2017.

[17] S. Bakas et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," *arXiv preprint arXiv:1811.02629*, 2018. doi: 10.48550/arXiv.1811.02629.

[18] P. Chang et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *American Journal of Neuroradiology*, 39(7):1201–1207, 2018.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[20] C. Ge, I. Y.-H. Gu, A. S. Jakola, J. Yang. Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging*, 20(1):87, 2020.

[21] F. Isensee et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[22] S. M. Lundberg, S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[23] B. H. Menze et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[24] M. T. Ribeiro, S. Singh, C. Guestrin. "Why should I trust you?": explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[25] A. Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[26] S. Bakas et al. Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, 2017.

[27] H. Verma, K. Im, P. Maji, A. Gupta. An intuitionistic fuzzy logic driven UNet architecture: application to brain image segmentation (IF-UNet). *arXiv preprint arXiv:2602.04227*, 2026. doi: 10.48550/arXiv.2602.04227.

[28] J. Arora et al. Conditional spatial intuitionistic fuzzy clustering for brain MRI. *Frontiers in Computational Neuroscience*, 18, 2024. doi: 10.3389/fncom.2024.1425008.

Appendix A Supplementary material

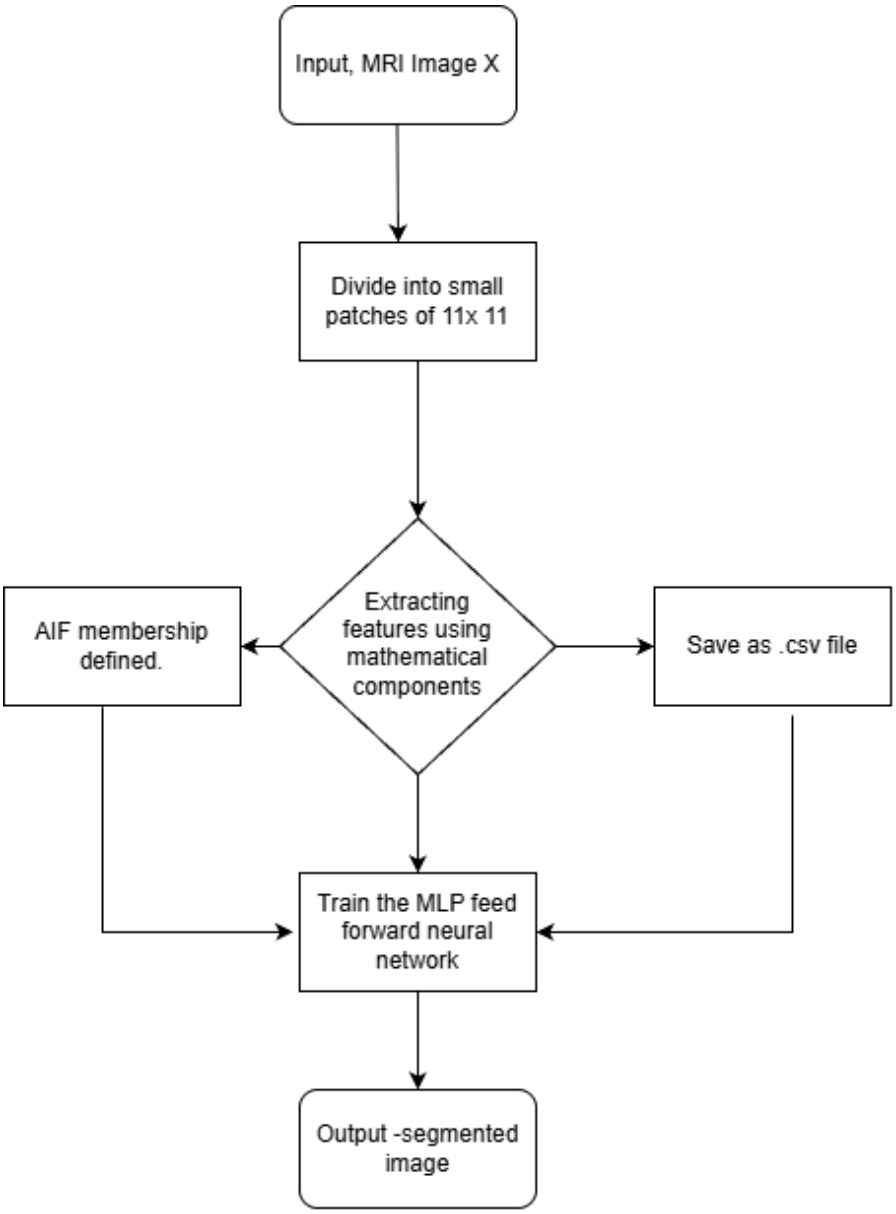


Fig. 1 Workflow of the classifier: a FLAIR volume is z-scored inside the brain mask, divided into 11×11 patches, described by five features, converted into membership, non-membership and hesitation degrees by the AIFS layer, and classified by a small MLP, which also returns the hesitation.

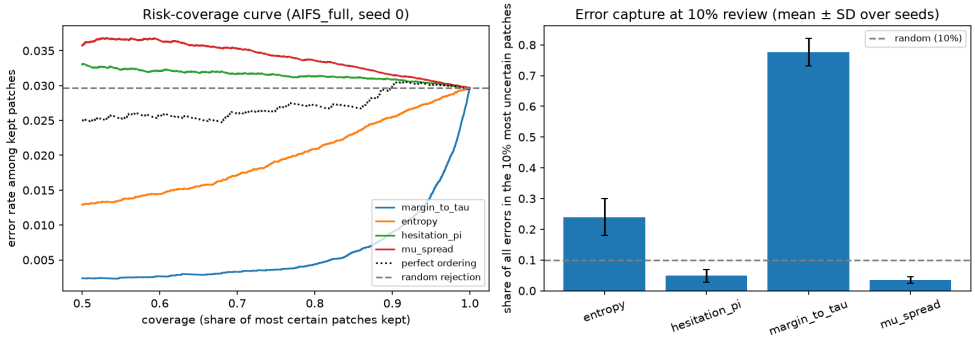


Fig. 2 Left: risk-coverage curves of four uncertainty scores for the AIFS model (seed 0). Right: share of all errors found among the 10% most uncertain patches (mean ± SD over five seeds); the dashed line marks random ordering.

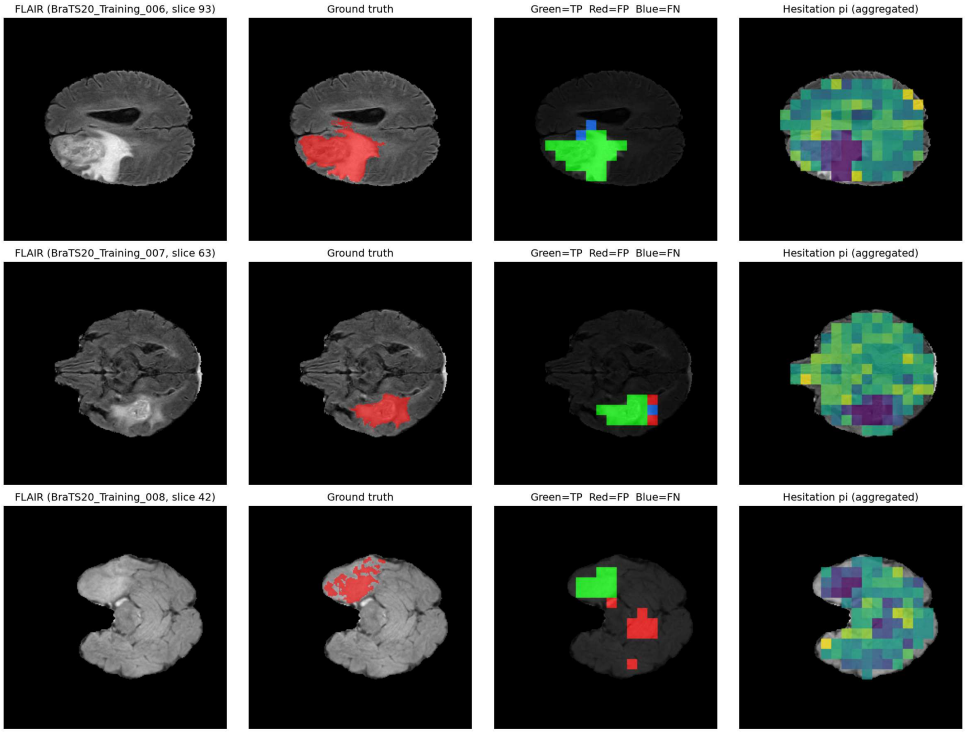


Fig. 3 Qualitative results for three unseen test patients: FLAIR slice, ground truth, overlay of true positives (green), false positives (red) and false negatives (blue), and map of the aggregated hesitation.

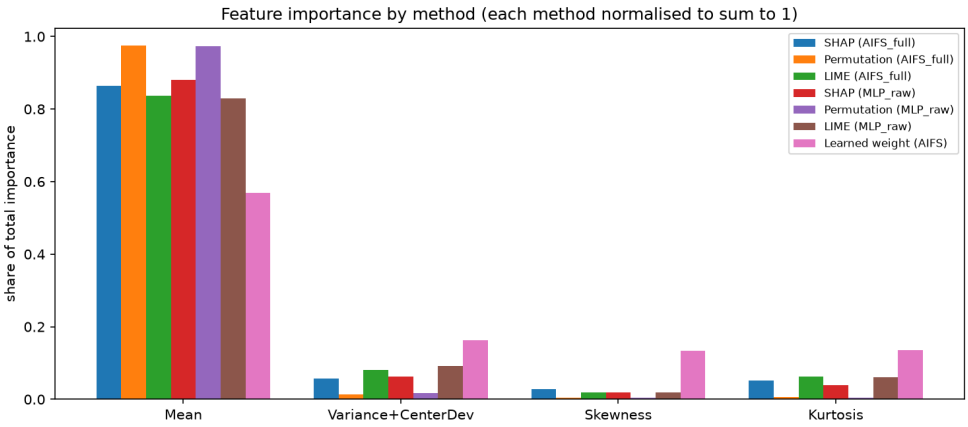


Fig. 4 Importance of the four feature groups by SHAP, permutation importance, LIME and the learned weights, for the AIFS and the MLP models (each method normalized to sum to 1; mean over three seeds).

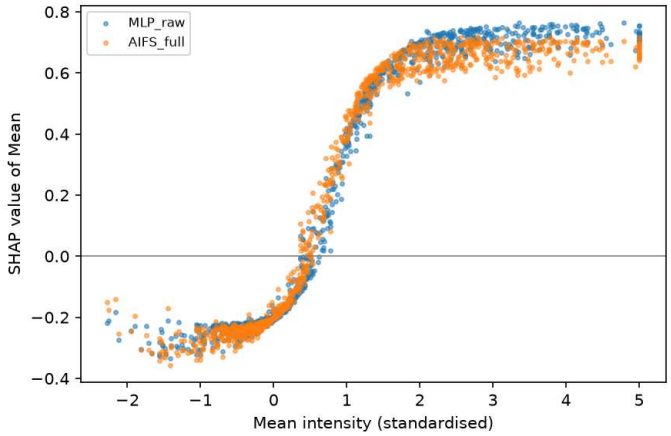


Fig. 5 SHAP value of the mean intensity against its standardized value for the MLP and the AIFS model (500 tumor and 500 non-tumor test patches, seed 0).

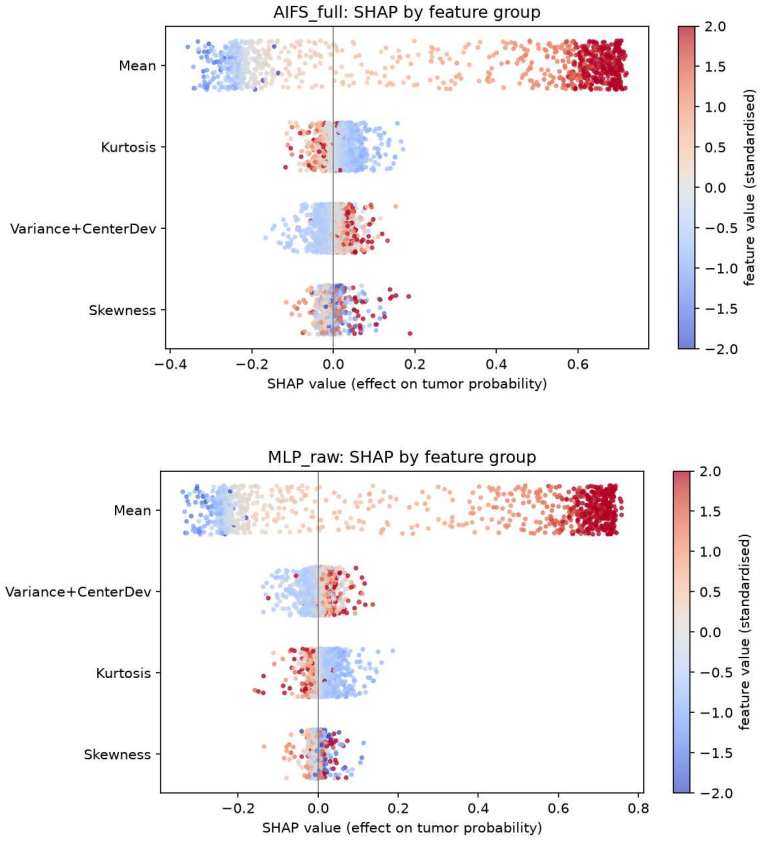


Fig. S1 Beeswarm plots of SHAP values for the AIFS model (top) and the MLP (bottom); colour shows the standardized feature value.

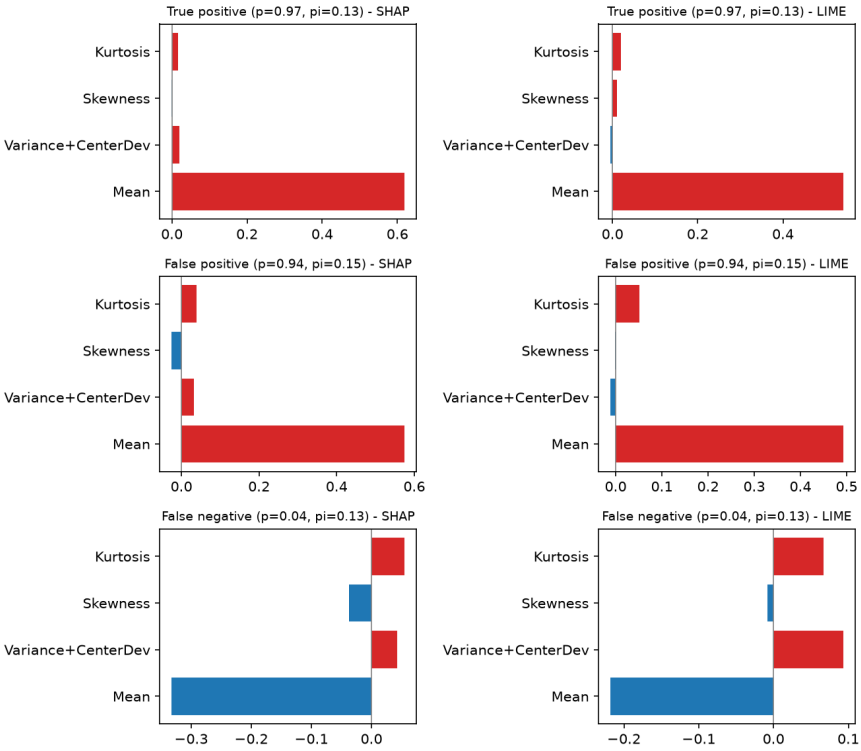


Fig. S2 Local SHAP and LIME explanations of one true positive, one false positive and one false negative.