ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

NEUROCALIB: Design, Implementation and Internal Evaluation of a Calibrated, Explainable and Selective Bimodular Framework for Brain Tumor Classification and Segmentation on MRI

Hervé Kinkete Mfumabi[1]; Matangila Gradi[2]; Nsiala Ndona Clémence[2]; Kavugho Kahasa Thérèse[2]; Kalema Manzambie Jordanie[2]; Kema Mulumba Joël[2]; Madio Motemona Godard[2]; Ilolo Nana[2]

[1]Faculty Academic Secretary, Faculty of Computer Science and Artificial Intelligence, William Booth University, Kinshasa, Democratic Republic of the Congo
[2]Students, Faculty of Computer Science and Artificial Intelligence, William Booth University, Kinshasa, Democratic Republic of the Congo

Publication Date: 2026/08/31

Abstract: High aggregate performance does not guarantee reliable individual predictions or clinical transportability. This study designed, implemented, and internally evaluated NEUROCALIB, a selective bimodular research framework comprising independently assessed classification and segmentation pipelines, and audited their integration into the NeuroVision web prototype. EfficientNetB0 was evaluated on 6,683 public brain MRI images across four classes; MobileNetV2–U-Net was evaluated on 4,048 image-mask pairs after removal of 189 exact duplicates. Validation data were used for model selection, temperature scaling, and abstention thresholds before the test sets were opened. On 1,003 classification test images, the verified V4 execution achieved 94.42% accuracy (95% CI, 93.02–95.81), 94.22% macro-F1 (92.68–95.61), and 99.57% macro one-versus-rest AUROC. Temperature scaling reduced ECE from 2.30% to 1.88%. At an entropy threshold of 0.333030, selective accuracy reached 99.00% at 89.93% coverage (902 accepted; 101 abstained). On 404 segmentation test images, deterministic mean Dice was 91.86%, eight-pass Monte Carlo averaging yielded 92.26%, and selective Dice reached 93.53% on 361 accepted images (89.36% coverage); the selection-specific gain was therefore 1.27 percentage points over Monte Carlo averaging. CPU benchmarks were 18.13 ms/image for classification and 436.37 ms/image for segmentation; eight Monte Carlo passes over 404 images required 1,364.61 s. The audited FastAPI branch aligned preprocessing, temperatures, mask and uncertainty thresholds, and refusal behavior, but Git LFS pointers prevented executable end-to-end verification with the definitive weights. NEUROCALIB provides image-level internal algorithmic evidence and an auditable engineering prototype, not patient-level, external, prospective, or clinical validation.

Keywords: *Brain MRI; Brain Tumor; Calibration; EfficientNetB0; Explainable AI; FastAPI; Selective Prediction; Segmentation; Uncertainty.*

How to Cite: Hervé Kinkete Mfumabi; Matangila Gradi; Nsiala Ndona Clémence; Kavugho Kahasa Thérèse; Kalema Manzambie Jordanie; Kema Mulumba Joël; Madio Motemona Godard; Ilolo Nana (2026) NEUROCALIB: Design, Implementation and Internal Evaluation of a Calibrated, Explainable and Selective Bimodular Framework for Brain Tumor Classification and Segmentation on MRI. *International Journal of Innovative Science and Research Technology*, 11(8), 2098-2110. https://doi.org/10.38124/ijisrt/26aug907

I. INTRODUCTION

Brain tumors display substantial variation in location, tissue contrast, infiltration, and acquisition characteristics. MRI is central to lesion detection and follow-up, yet expert interpretation and delineation remain labor-intensive. Deep learning can support these tasks, but a high mean score alone does not establish reliable case-level behavior. Dataset leakage, poorly calibrated probabilities, selective reporting, and untested software integration can produce apparently strong systems that fail under clinically relevant variation [1], [2], [28], [32].

Classification and segmentation answer complementary questions. EfficientNet offers a favorable capacity-to-cost trade-off for transfer learning [11], whereas U-Net and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

MobileNetV2 provide an efficient encoder-decoder basis for pixel-level delineation [12], [13]. However, combining two architectures in one manuscript does not demonstrate a clinical sequence unless both operate on paired images from the same patients. NEUROCALIB therefore uses the term bimodular framework: the classifier and segmenter were developed and evaluated independently on two unpaired public datasets.

Reliability was operationalized through post-hoc temperature scaling [15], Monte Carlo dropout [16], uncertainty-based failure detection [24], [29], and selective prediction [17], [25]. Grad-CAM was included as a qualitative audit aid [14], not as anatomical segmentation or causal evidence. The central research question was: can two moderate-cost architectures provide accurate internal predictions while identifying outputs that should be referred for human review? Two hypotheses were prespecified: temperature scaling should reduce held-out calibration error without changing the predicted class, and validation-derived uncertainty thresholds should reduce residual error among accepted cases while reporting the associated coverage.

The contribution is a controlled integration of six elements: duplicate auditing, independent module evaluation, post-hoc calibration, uncertainty estimation, selective abstention, and auditable software orchestration. No superiority over deep ensembles, conformal predictors, or alternative selective baselines is claimed because those methods were not compared on identical partitions. The study follows CLAIM 2024 and the corrected STARD-AI reporting principles while remaining explicitly an internal, image-level evaluation [20], [21], [31].

II. RELATED WORK AND RESEARCH GAP

Recent brain-tumor classifiers often report accuracies above 95%, but direct comparison is unsafe when datasets, deduplication rules, statistical units, augmentation, and test construction differ [3]–[5], [7], [9]. Segmentation studies increasingly use attention, multi-scale blocks, ASPP, and multimodal 3-D inputs [1], [6], [8], [10]. These approaches may improve representation but require more computation and commonly address datasets and target definitions that differ from the present 2-D public corpus.

Uncertainty work in medical imaging emphasizes that calibration, discrimination, and failure detection are distinct properties [2], [23]–[25], [29], [30]. Saliency maps additionally require human-centered validation and stability testing before clinical interpretation [26], [27]. The research gap addressed here is therefore not a new convolutional block; it is a reproducible, resource-aware protocol that links performance, calibration, abstention, explicit failure accounting, and prototype audit while separating the evidence produced by each module.

Table 1 Positioning Against Representative Work

Study	Primary focus	Key distinction
Iqbal et al. [4]	EfficientNetB0 classification	Different dataset/protocol
Vimala et al. [5]	Hybrid classification	No shared split
Preetha et al. [8]	Attention U-Net segmentation	Different target definition
Mehrtash et al. [23]	Segmentation calibration	Uncertainty-centered
NEUROCALIB	Two independent selective modules	Calibration, abstention, audit

III. MATERIALS AND METHODS

➢ *Study Design, Data, and Statistical Units*

A retrospective internal-development design was used. Training, validation, and testing were separated; epoch selection, temperatures, binarization choices, uncertainty thresholds, and target coverage were determined without optimizing on the test sets. The classification dataset was a cleaned local copy of the public Brain Tumor MRI Dataset [18]. After audit, 6,683 images remained: 4,678 training, 1,002 validation, and 1,003 testing images. The four labels were glioma, meningioma, no tumor, and pituitary tumor.

The segmentation dataset was the public Brain Tumor Segmentation Dataset [19]. Pairing identified 4,237 image-mask pairs; 189 exact duplicate images were removed, leaving 4,048 pairs split 3,240/404/404. The numeric labels 0–3 were retained because a robust clinical label dictionary was unavailable. Class 0 included 1,422 empty masks. Exact-hash auditing reduces one leakage mechanism but cannot identify adjacent slices, reconstructions, or derived variants from the same subject.

Neither source provided dependable patient identifiers. Consequently, patient-level separation could not be guaranteed, and images—not patients—were the experimental and bootstrap units. Confidence intervals are conditional on these files and may be too narrow if within-patient correlation exists. This limitation cannot be removed by wording and requires a patient-identifiable corpus.

Table 2 Dataset Composition After Audit

Task / class	Train	Validation	Test	Total
Classification	4678	1002	1003	6683
Glioma	1245	267	267	1779
Meningioma	1042	224	223	1489
No tumor	1160	248	249	1657
Pituitary	1231	263	264	1758

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

Segmentation pairs	3240	404	404	4048
Class 0	1138	142	142	1422
Class 1	519	65	65	649
Class 2	795	99	99	993
Class 3	788	98	98	984

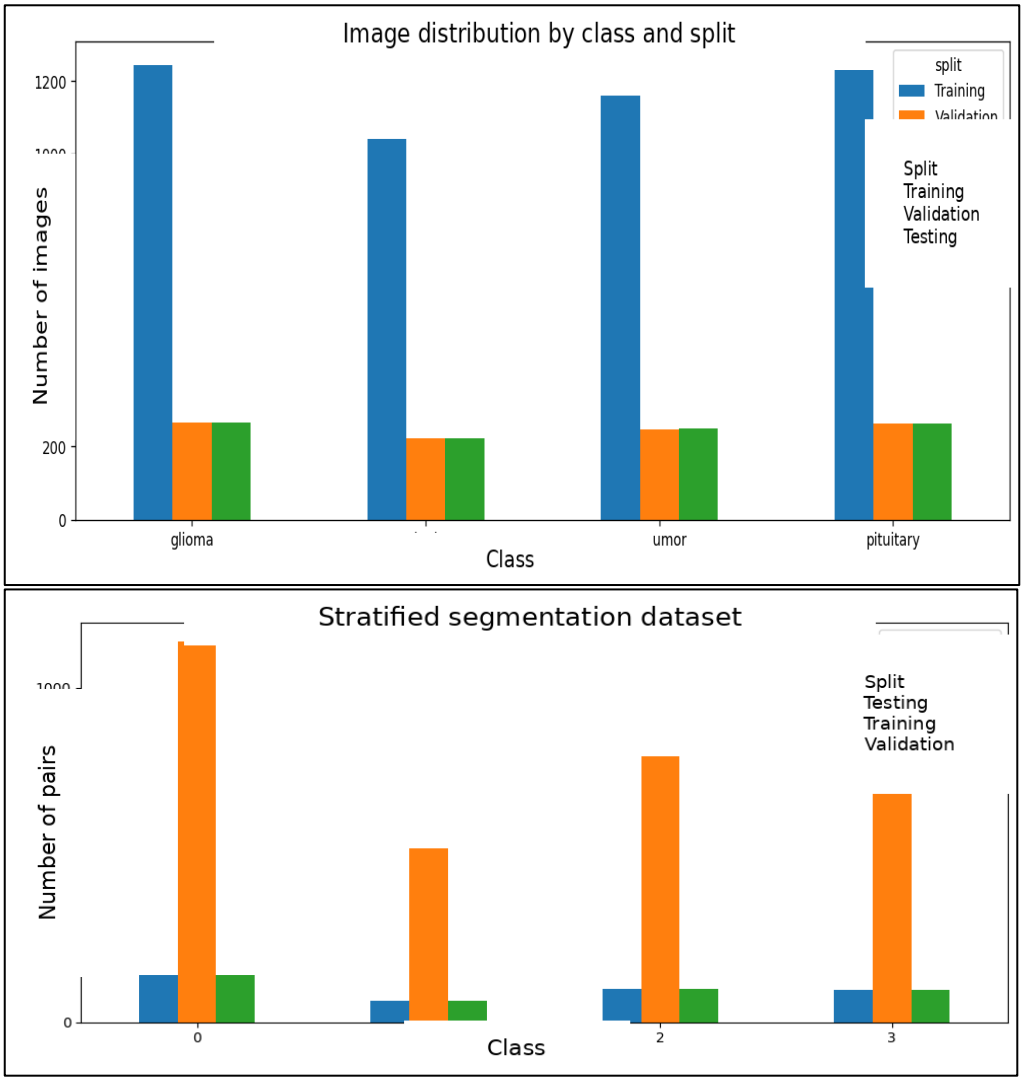

Fig 1 Distribution of Experimental Units by Class and Split: Classification (Left) and Stratified Segmentation (Right).

➢ *Preprocessing and Augmentation*

Classification images were decoded as three channels, resized to 224×224 pixels, and organized through Keras utilities. Training-only stochastic augmentation was used; validation and test images received deterministic preprocessing. A reproducibility supplement records a 7,200-file inventory, 322 redundant copies excluded during corpus curation, 195 offline-augmented images excluded from final splits, final manifests, and source hashes.

Segmentation RGB images and masks were resized to 192×192 pixels. Image intensities were normalized according to the MobileNetV2 input convention and masks were binarized. Synchronized spatial transformations were applied to image-mask pairs during training only. The test set and all validation-based threshold selection remained free from stochastic augmentation.

➢ *Classification Architecture and Training*

ImageNet-pretrained EfficientNetB0 served as a frozen feature extractor. A dense classification head used global average pooling, batch normalization, a 256-unit layer, dropout of 0.30, and a four-class softmax output. Adam with an initial learning rate of 10^{-3} optimized categorical cross-entropy. Early stopping and checkpointing used validation loss. The definitive V4 checkpoint contained approximately 4.42 million parameters and occupied 20.49 MB.

The best checkpoint occurred at epoch 15. Training accuracy reached 100.00% and validation accuracy 95.71%, an observed gap of 4.29 percentage points. This overfitting signal is reported rather than concealed and motivates future fine-tuning and regularization comparisons.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

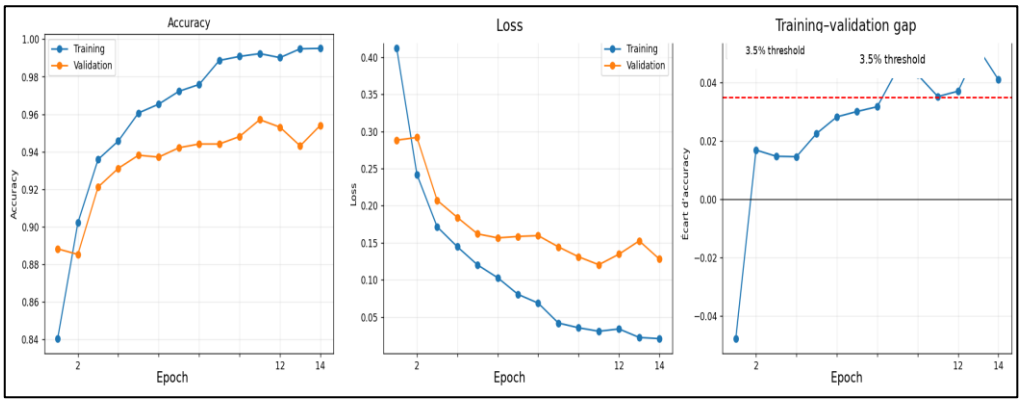

Fig 2 EfficientNetB0 Learning Curves for Accuracy, Loss, and Macro-F1 in the Verified V4 Notebook.

➢ *Segmentation Architecture and Training*

The segmenter was a U-Net decoder with a MobileNetV2 encoder. Four decoder stages combined upsampling, batch normalization, spatial dropout, convolution, and encoder skip connections. The final sigmoid layer generated a tumor probability map. The network contained 3,029,537 parameters and occupied 32.21 MB.

Training used a warm-up phase followed by partial fine-tuning. Adam used 10^{-3} initially and 10^{-5} during fine-tuning. The loss combined Dice loss and focal loss (0.60/0.40), with focal α=0.75 and γ=2. The output was binarized at the validation-selected threshold of 0.35. Batch-normalization layers remained non-trainable during fine-tuning to avoid unstable test-dependent behavior.

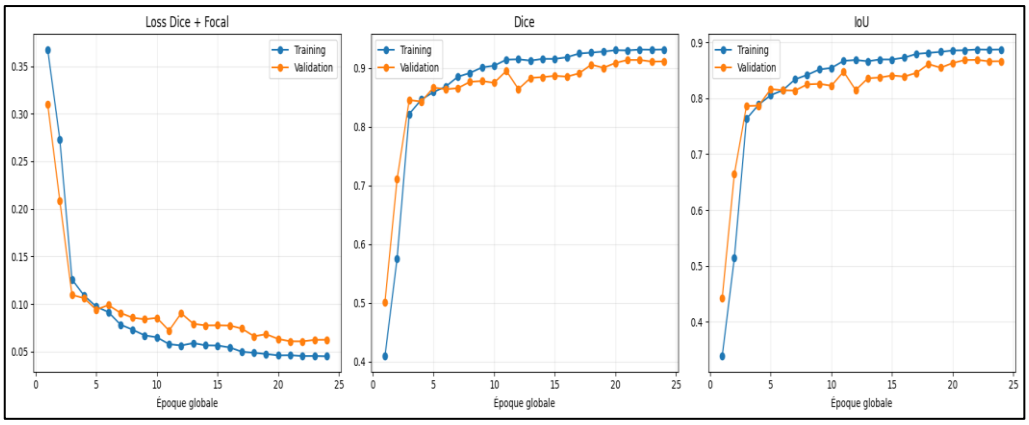

Fig 3 Loss, Dice, and IoU During Warm-up and Partial Fine-Tuning of MobileNetV2–U-Net.

➢ *Calibration, Uncertainty, and Selective Prediction*

Temperature scaling divided classification logits by a positive scalar and applied the corresponding transformation to segmentation logits [15]. Temperatures were optimized on validation data and then frozen: T=1.223167 for classification and T=1.973524 for segmentation. Calibration was summarized by negative log-likelihood (NLL), Brier score, and expected calibration error (ECE) using 15 equal-width bins. Segmentation ECE over all pixels is considered exploratory because background pixels dominate; foreground and boundary calibration require archived pixel outputs and are prespecified for replication.

Classification abstention used normalized predictive entropy. The threshold 0.333030 was the validation-derived rule targeting approximately 90% coverage. A secondary Monte Carlo diagnostic used 30 passes with the EfficientNet backbone held in inference mode and only a post-hoc dropout layer activated at rate 0.05. This isolates stochasticity from batch-normalization statistics.

Segmentation used eight stochastic passes. The encoder and batch-normalization layers remained in inference mode, while decoder dropout provided variation. The mean probability map was the Monte Carlo prediction and pixel variance represented uncertainty. Images with mean uncertainty above 0.00566, fixed on validation, were referred for review.

➢ *Robustness, Explainability, and Statistics*

The reproducible classification stress test added Gaussian noise with mean 0 and standard deviation 45 on the 0–255 input scale to all 1,002 validation images, followed by clipping. It quantifies one synthetic corruption only and is not an external out-of-distribution test. Grad-CAM maps were generated for qualitative inspection [14]; no radiologist ratings, independent localization annotations, stability repetitions, or quantitative localization analysis were performed.

Classification metrics included accuracy, macro precision, recall, F1, specificity, one-versus-rest AUROC, average precision, and confusion matrices. Segmentation

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

metrics included mean image Dice, global Dice, IoU, precision, recall, specificity, pixel accuracy, and critical empty/non-empty mask errors. Ninety-five-percent confidence intervals used 1,000 image-level bootstrap replications. Selective results always report the full-set metric, accepted-set metric, coverage, and number abstained.

➢ *Software Architecture and Reproducibility*

NeuroVision uses a browser client and a FastAPI backend. The aligned /api/predict design performs file validation, module-specific preprocessing, independent model calls, calibration, uncertainty computation, abstention, and JSON response construction. Classification uses 224×224 input, T=1.223167, and entropy threshold 0.333030. Segmentation uses 192×192 input, T=1.973524, eight Monte Carlo passes, mask threshold 0.35, and uncertainty threshold 0.00566.

The audited repository contained 133-byte Git LFS pointer files rather than executable HDF5 weights. The segmenter pointer announced 93,530,416 bytes, inconsistent with the 32.21-MB definitive notebook artifact. Therefore, source-level alignment and three static tests were verified, but API-notebook numerical parity, deployment latency, concurrency, and end-to-end success were not measured. The corrected code rejects inference with HTTP 503 when verified weights are unavailable and contains no silent fallback to synthetic models.

IV. RESULTS

➢ *Classification Performance*

On 1,003 test images, the verified V4 execution produced 947 correct and 56 incorrect decisions: accuracy 94.42% (95% CI, 93.02–95.81), macro-F1 94.22% (92.68–95.61), macro specificity 98.15%, and macro one-versus-rest AUROC 99.57%. Meningioma remained the most difficult class with 88.49% F1. The glioma–meningioma pair accounted for 29 of 56 errors (15 gliomas predicted as meningiomas and 14 meningiomas predicted as gliomas). Five tumor images were predicted as no tumor and seven no-tumor images as tumor classes.

Table 3 Classification Test Performance

Class / level	Precision	Recall	F1	Specificity	AUROC
Macro	94.25%	94.20%	94.22%	98.15%	99.57%
Glioma	93.66%	94.01%	93.83%	97.69%	99.41%
Meningioma	89.09%	87.89%	88.49%	96.92%	99.10%
No tumor	97.98%	97.19%	97.58%	99.34%	99.93%
Pituitary	96.27%	97.73%	96.99%	98.65%	99.83%

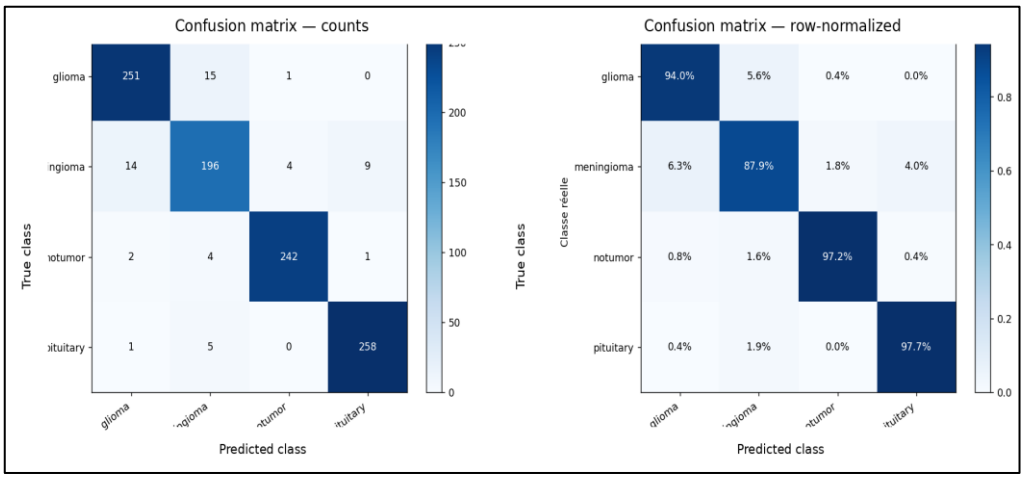

Fig 4 Raw and Row-Normalized Confusion Matrices for the 1,003-Image Classification Test Set.

➢ *Classification Calibration and Selectivity*

Temperature scaling changed neither argmax predictions nor the 56 deterministic errors. Test NLL decreased from 0.153715 to 0.144511, Brier score from 0.078045 to 0.075433, and ECE from 2.3035% to 1.8786%. The validation-derived entropy threshold accepted 902 of 1,003 images and abstained on 101. Accepted-case accuracy was 99.00% and macro-F1 98.90% at 89.93% coverage. These selective values do not describe the entire test set and must be read with the 10.07% abstention rate.

Table 4 Calibration and Selective Classification

Indicator	Raw / full	Calibrated / accepted	Interpretation
NLL	0.153715	0.144511	−5.99%
Brier	0.078045	0.075433	−3.35%
ECE	2.3035%	1.8786%	−0.4249 point
Accuracy	94.42%	99.00%	89.93% coverage
Macro-F1	94.22%	98.90%	101 abstentions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

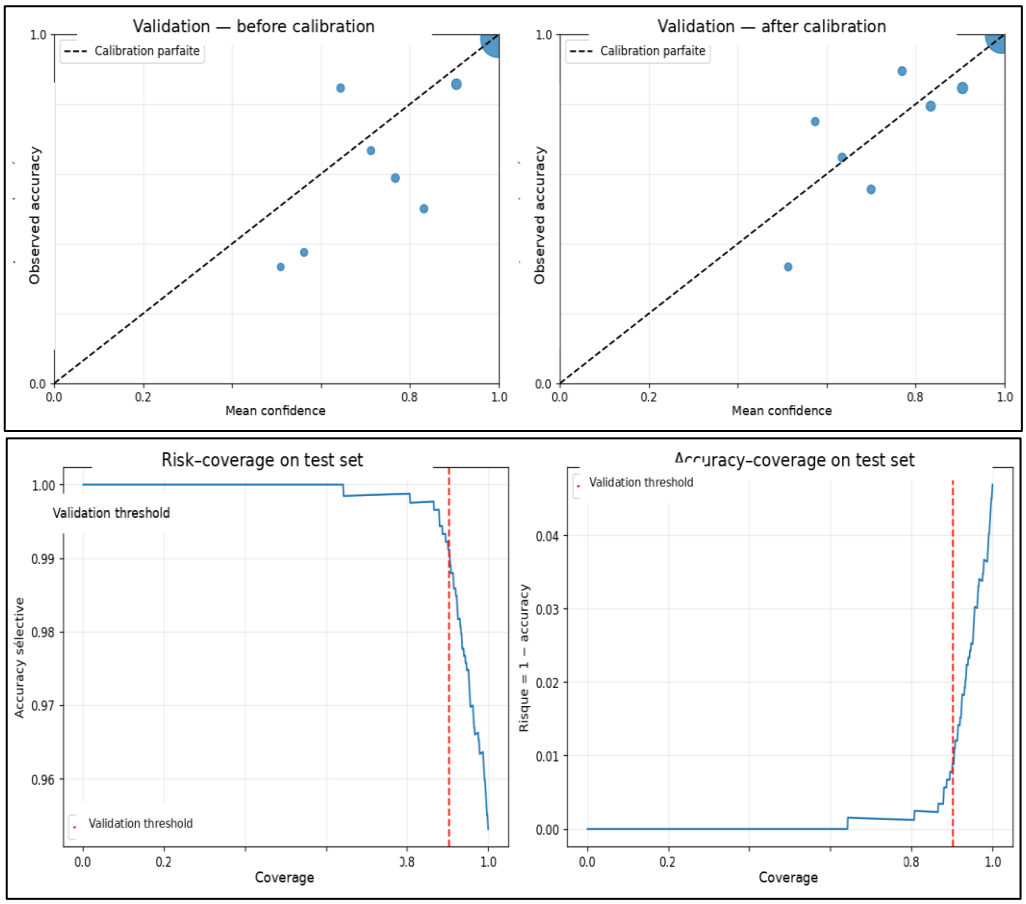

Fig 5 Classification Calibration Before/After Temperature Scaling (Left) and Risk–Coverage Behavior on Testing (Right).

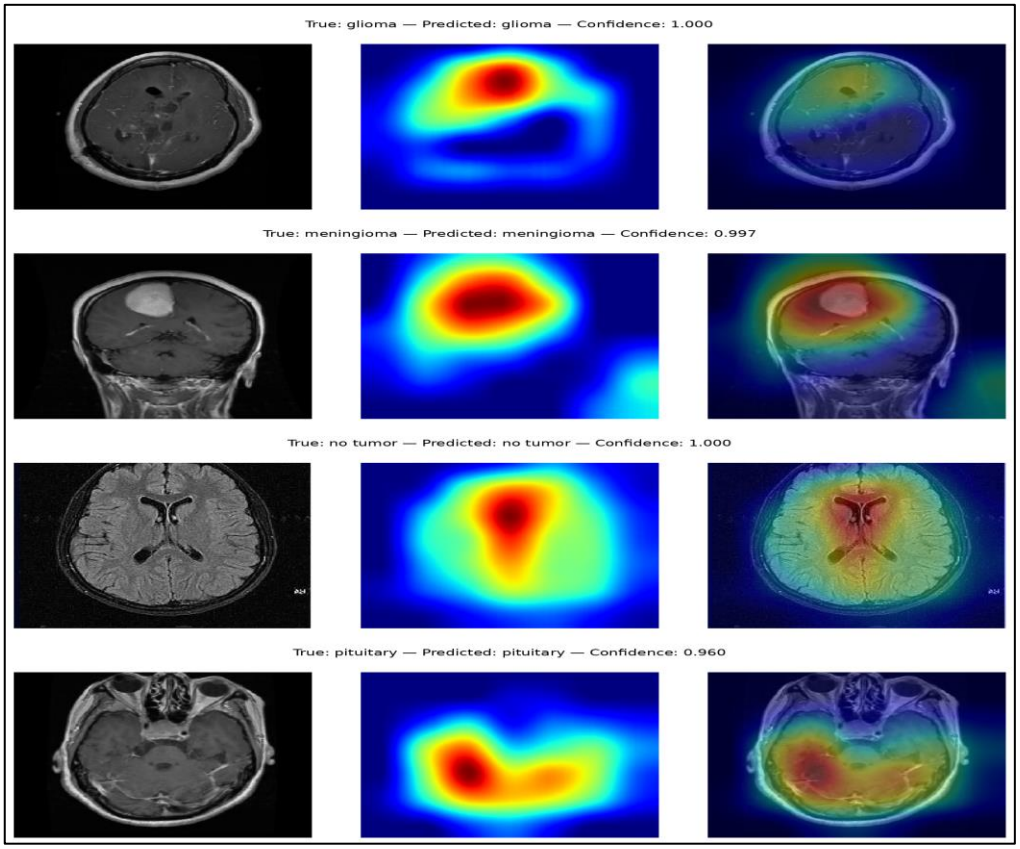

Fig 6 Representative Grad-CAM Outputs from the Verified V4 Notebook. These Qualitative Maps are Neither Anatomical Masks NOR Clinical Validation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

The isolated Monte Carlo diagnostic showed negligible deterministic drift ($<10^{-7}$). Mean predictive entropy was 0.120272 on testing and mean mutual information 0.025490. The Gaussian-noise stress test yielded an AUROC of 0.7717 for clean-versus-corrupted discrimination, indicating moderate sensitivity to this specific corruption rather than general OOD capability.

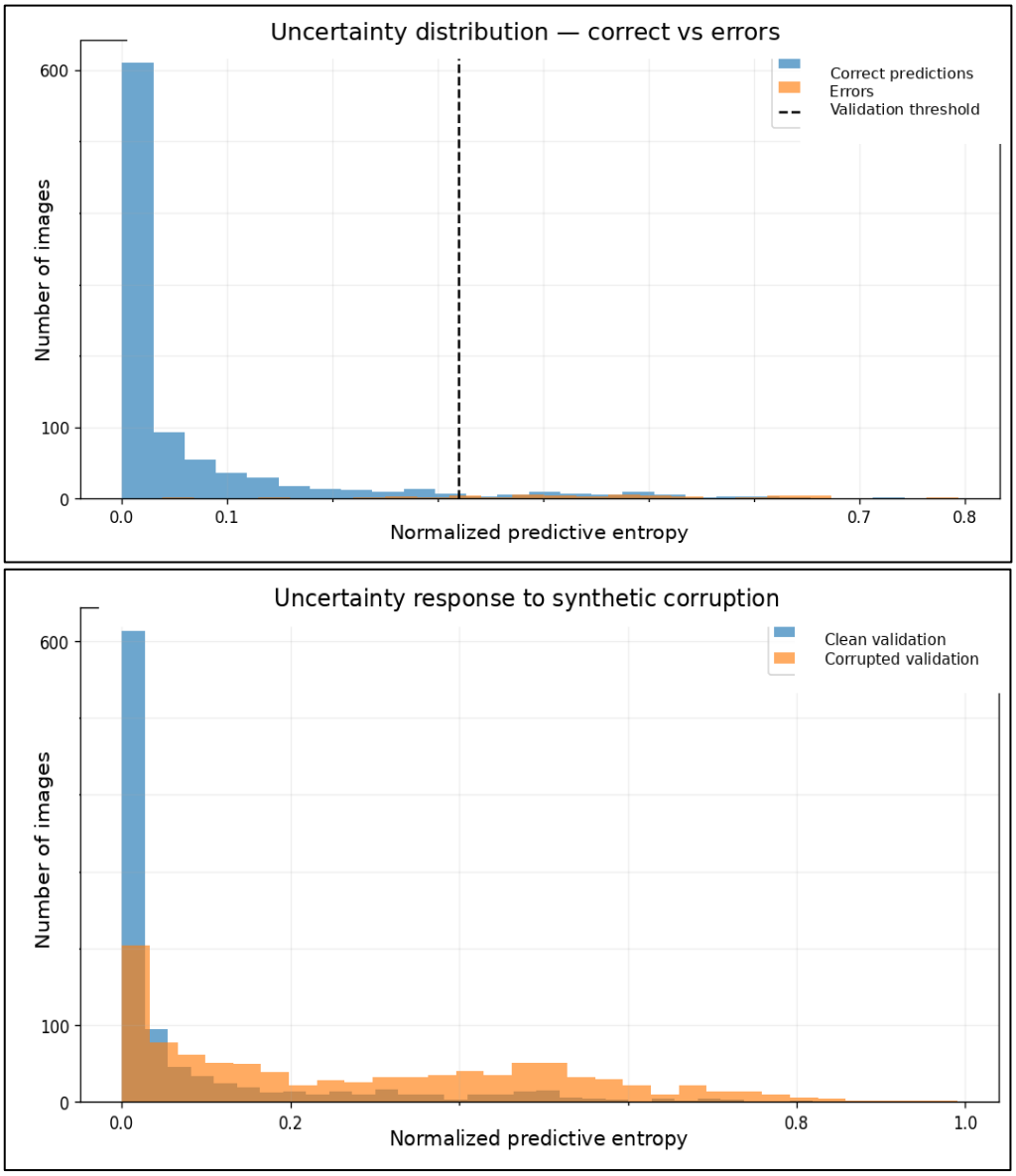

Fig 7 Uncertainty-Based Error Separation (Left) and Response to the Prespecified Gaussian Corruption (Right).

➢ *Segmentation Performance*

Across 404 test images, deterministic mean image Dice was 91.86% (95% CI, 90.10–93.48) and mean IoU 87.71% (85.66–89.45). Mean precision was 93.55%, recall 92.78%, specificity 99.91%, and pixel accuracy 99.76%. Global Dice from aggregated pixels was 89.73%. On the 262 tumor-positive images, tumor Dice was 87.84%. Class 1 was the weakest subgroup at 76.64% Dice, whereas class 0 reached 99.30% because empty masks were comparatively easy.

Table 5 Segmentation Test Performance

Group	n	Dice	IoU	Precision	Recall
All images	404	91.86%	87.71%	93.55%	92.78%
Tumor positive	262	87.84%	81.42%	90.49%	88.87%
Class 0	142	99.30%	99.30%	99.30%	100.00%
Class 1	65	76.64%	67.44%	82.03%	79.26%
Class 2	99	91.00%	85.94%	94.77%	90.35%
Class 3	98	92.07%	86.14%	91.79%	93.74%

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

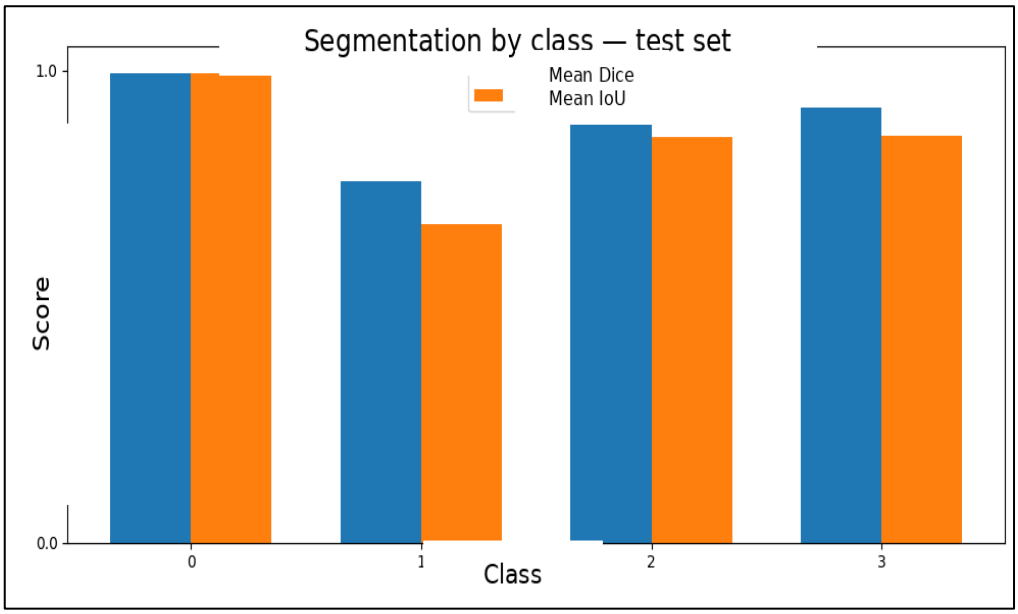

Fig 8 Mean Dice and IoU by Numeric Segmentation Class on Testing.

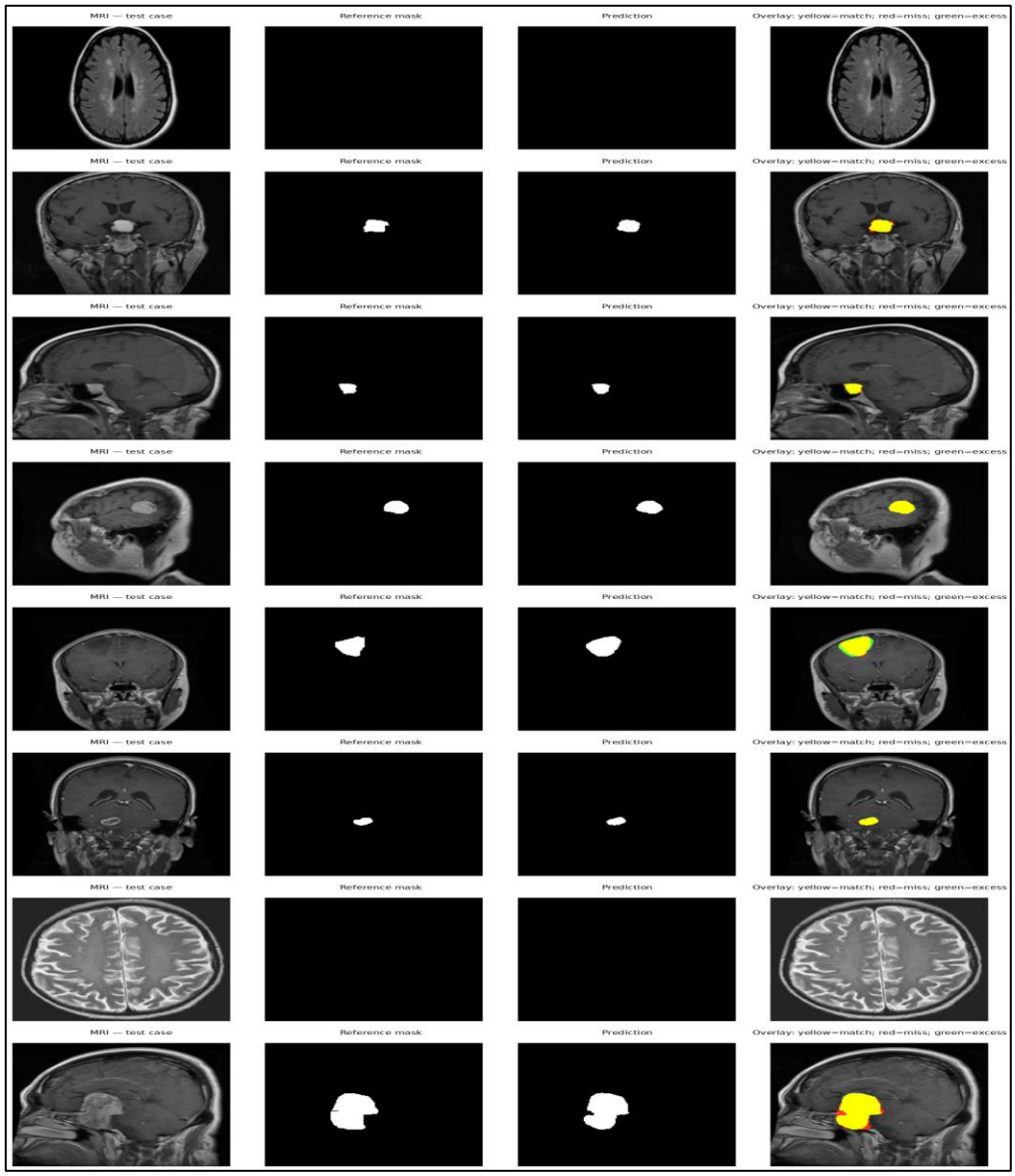

Fig 9 Qualitative Segmentation Examples: MRI, Reference Mask, Probability Map, and Predicted Overlay.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

➢ *Segmentation Calibration and Selectivity*

Temperature scaling reduced global pixel-calibration metrics (validation NLL 0.01743→0.00987 and ECE 0.199%→0.082%; testing NLL 0.01607→0.00968, Brier 0.00235→0.00214, and ECE 0.223%→0.126%). Because these estimates are dominated by background, they do not establish tumor or boundary calibration; no unarchived foreground value was manufactured.

The selective analysis distinguishes three stages. Deterministic inference achieved 91.86% Dice on all 404 images. Eight-pass Monte Carlo averaging, before abstention, achieved 92.26% on the same 404 images, a 0.40-point gain. Applying the frozen uncertainty threshold accepted 361 images and referred 43, yielding 93.53% Dice at 89.36% coverage. The gain specifically attributable to selection is therefore 93.53−92.26=1.27 points, not 1.67 points relative to deterministic inference.

Table 6 Deterministic, Monte Carlo, and Selective Segmentation

Mode	Evaluated	Coverage	Mean Dice	Attribution
Deterministic	404	100.00%	91.86%	Reference
8-pass MC	404	100.00%	92.26%	+0.40 vs deterministic
MC + abstention	361 accepted	89.36%	93.53%	+1.27 vs MC

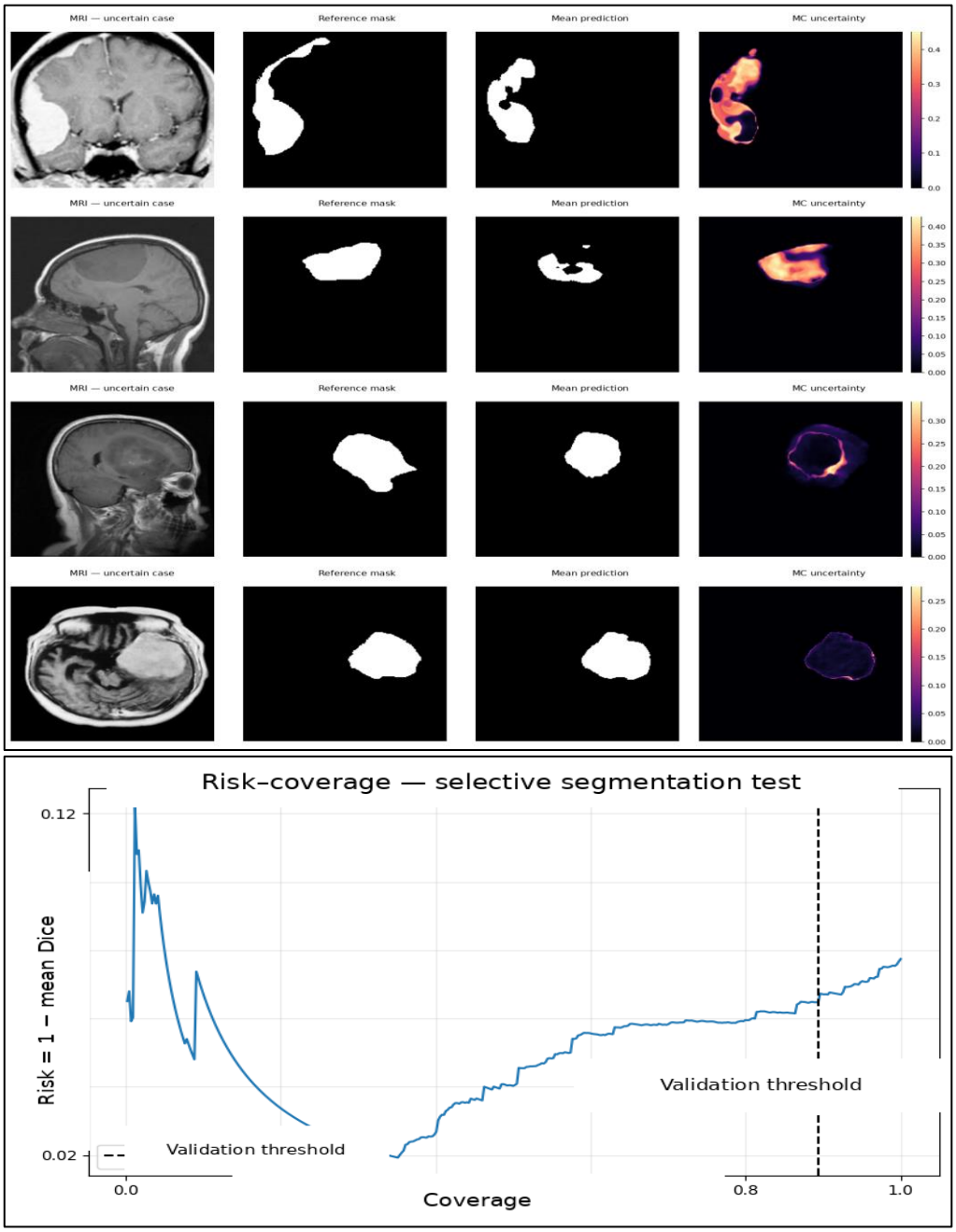

Fig 10 Uncertain Segmentation Examples and Variance Maps (Left); Segmentation Risk–Coverage Curve with the Frozen Operating Point (Right).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

➢ *Computational and Software Results*

Component benchmarks measured 18.13 ms/image for classification on 160 CPU images and 436.37 ms/image for segmentation on 80 CPU images. Eight Monte Carlo passes over 404 images required 1,364.61 s, or 3.38 s/image for the complete eight-pass procedure. These measurements describe notebook components in different execution environments and are not network end-to-end Render latency.

Table 7 Computational and Software Evidence

Component	Sample	Result	Evidence boundary
Classification CPU	160 images	2.90 s; 18.13 ms/image	V4 notebook
Segmentation CPU	80 images	34.91 s; 436.37 ms/image	Segmentation notebook
8-pass MC segmentation	404×8 passes	1,364.61 s; 3.38 s/image	Eight-pass cost
Model size	2 files	20.49 MB; 32.21 MB	Saved artifacts
FastAPI audit	3 static tests	3/3 passed	No weight-based E2E

V. NEUROVISION PROTOTYPE: DESIGN, DEPLOYMENT, AND TECHNICAL VALIDATION

➢ *Bimodular Architecture*

NeuroVision operationalizes the engineering contribution through a browser interface, FastAPI service, independent preprocessing pipelines, two model calls, and a traceable JSON response. Displaying classification and segmentation together demonstrates software integration only; the datasets are unpaired, and no patient MRI was simultaneously classified and segmented in the experimental evaluation.

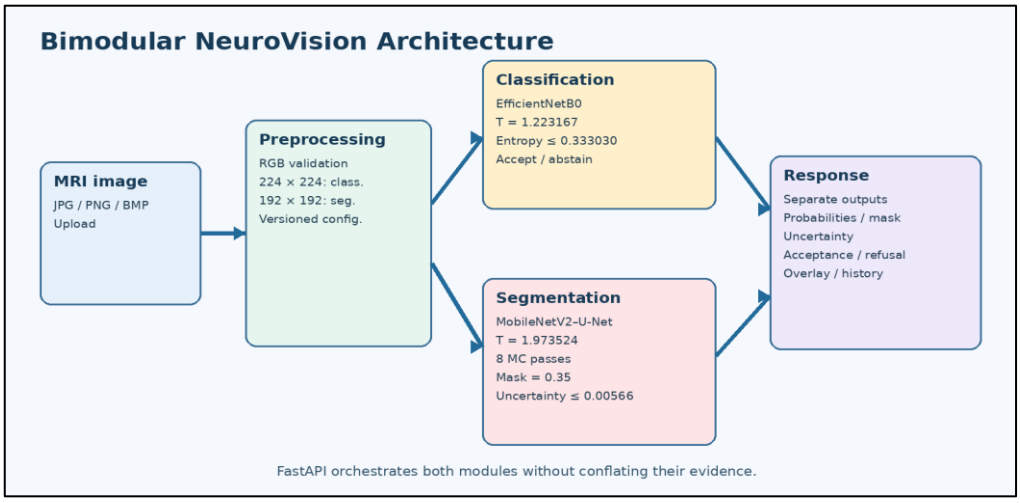

Fig 11 Audited Bimodular NeuroVision Architecture and Frozen Experimental Parameters.

➢ *API Alignment, Safety, and Traceability*

The corrected branch exposes module name, temperature, thresholds, uncertainty, and acceptance/abstention decision. The frontend performs no inference. Files are limited to 10 MB, invalid images are rejected, exceptions are handled, and missing scientific weights produce a degraded health state and HTTP 503 rather than fabricated outputs. A numerical parity test on a frozen image batch remains mandatory once definitive weights are available.

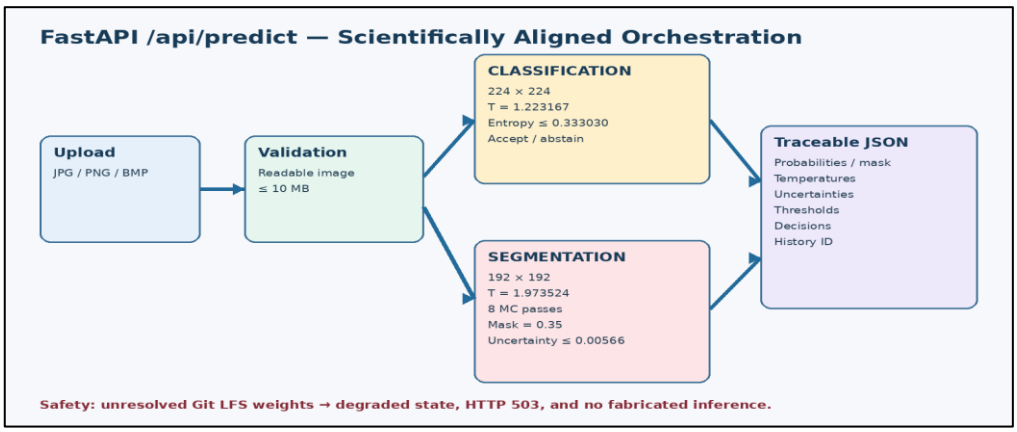

Fig 12 Aligned /Api/Predict Orchestration with Module-Specific Preprocessing, Calibration, Uncertainty, Abstention, and Explicit Refusal when Weights are Unavailable.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

➢ *User Interface and Non-Diagnostic Language*

The corrected interface labels the output as algorithmic classification requiring confirmation by a qualified professional. It presents calibrated probabilities, uncertainty, the acceptance/abstention decision, and segmentation output separately. Terms implying autonomous diagnosis and the fallback generation of fake model outputs were removed. The French-language screenshot reflects the localized prototype rather than the language of the manuscript.

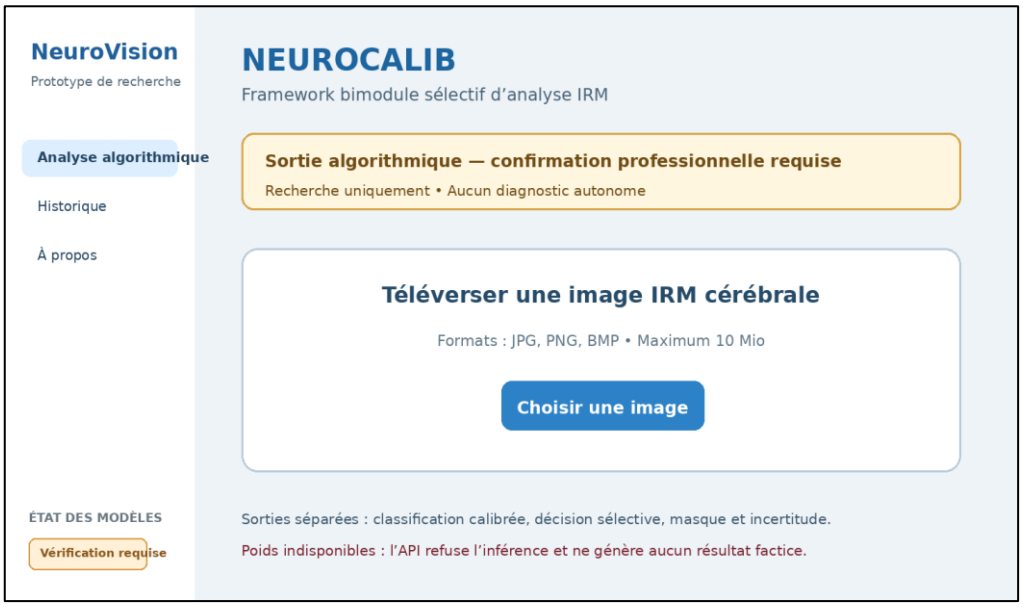

Fig 13 Localized NeuroVision Interface Mock-up with Non-Diagnostic Wording, Separate Outputs, and Explicit Model Status.

➢ *Render Deployment and Functional Protocol*

The repository defines a Render Web Service build and start chain. A five-step functional protocol was prespecified: interface and /api/health loading; valid upload and controlled invalid-input rejection; independent execution of both modules; rendering of probabilities, mask, uncertainty, and abstention; and history write/read. Source structure, syntax, constants, and negative refusal behavior were audited. Weight-dependent execution, p50/p95/p99 latency, memory, concurrency, restart persistence, and API-notebook parity remain not run and are reported as such.

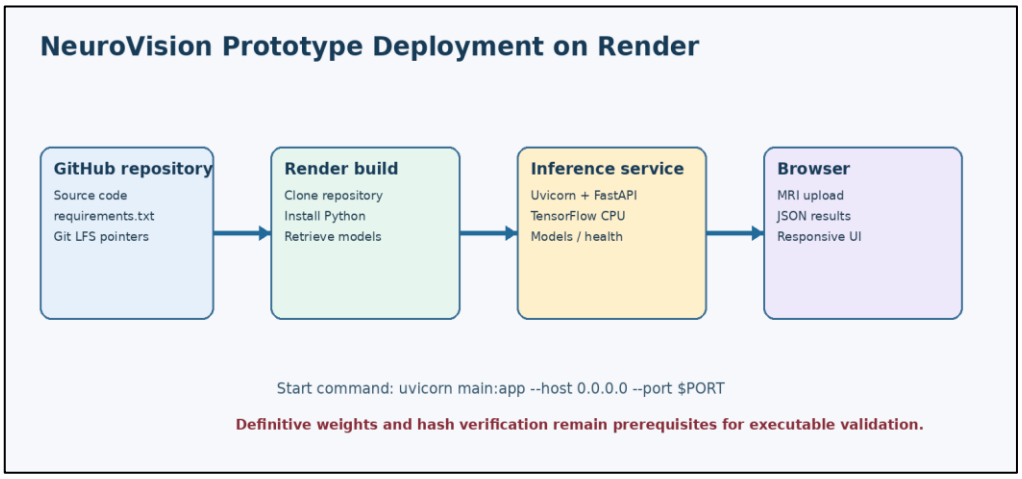

Fig 14 Intended Render Build and Execution Chain for the NeuroVision Demonstrator.

VI. DISCUSSIONS

The verified classification results demonstrate strong internal discrimination but not perfect generalization. The 4.29-point training-validation gap and glioma–meningioma concentration reveal where additional regularization, progressive fine-tuning, focal loss, or class-aware sampling should be evaluated. Comparison with published accuracies is descriptive because public datasets may include repeated subjects, derived slices, or incompatible partitions.

Calibration improved NLL, Brier score, and ECE without changing decisions, supporting H1 for this test set. Selective prediction materially reduced residual error among accepted cases, supporting H2 while preserving explicit coverage accounting. In classification, 99.00% accuracy applies to 902/1,003 images, not the complete test. In

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

segmentation, Monte Carlo averaging and abstention have distinct effects: +0.40 and +1.27 Dice points, respectively.

Segmentation performance was heterogeneous. The mean Dice of 91.86% and median of 96.67% indicate a tail of difficult cases, and class 1 Dice of 76.64% prevents an overly favorable interpretation. Global pixel accuracy and global ECE are background-sensitive. Future work should report tumor-only and boundary calibration, surface distance metrics, and clinically meaningful lesion-size subgroups.

Grad-CAM supports visual traceability only. Plausible heatmaps may accompany wrong decisions, and maps can be unstable without changing the class [26], [27]. A clinical XAI study should prespecify radiologist readers, localization targets, rating scales, inter-reader agreement, repeated explanations, and human-AI decision endpoints.

The software contribution is an audited implementation rather than an executed deployment validation. The inability to resolve Git LFS weights is a scientifically relevant negative finding: it prevents the service from being presented as a digital reproduction of the notebooks. The corrected refusal behavior is safer than a mock fallback, but real weights, immutable hashes, and parity tests are required before operational claims.

VII. LIMITATIONS AND PRESPECIFIED VALIDATION PATH

The main limitations are structural: (1) image-level rather than patient-level units; (2) no independent external cohort, center, scanner, or protocol; (3) unpaired classification and segmentation datasets; (4) a single fully specified synthetic corruption rather than genuine OOD evaluation; (5) background-dominated segmentation ECE; (6) qualitative Grad-CAM without reader evaluation; (7) incomplete public immutability of code, environments, and weights; and (8) no weight-based end-to-end Render test. Demographic metadata were unavailable, precluding fairness analysis by age, sex, origin, center, or scanner.

The evidence path is ordered: obtain patient identifiers and rebuild partitions and bootstrap intervals at patient level; freeze models and weights with hashes; evaluate an independent external cohort; compare uncertainty baselines on identical splits; quantify foreground and boundary calibration and multiple corruption severities; publish an immutable repository and DOI archive; execute API-notebook parity and end-to-end load tests; then conduct radiologist, fairness, human-AI interaction, drift-monitoring, prospective, and multicenter studies. Real DICOM workflows and governance must precede clinical deployment.

- *Validity Threats and Mitigation*

Internal validity requires patient identifiers and patient-level splits; statistical validity requires patient-level bootstrap and an independent cohort; construct validity requires foreground, boundary, and surface-distance metrics; external validity requires multicenter testing; clinical validity requires radiologist and prospective workflow studies; equity requires a governed cohort with demographic metadata; and software validity requires definitive hashes, API-notebook parity, latency, memory, and concurrency tests.

VIII. CONCLUSION

NEUROCALIB combines two independently evaluated algorithmic pipelines with an audited engineering prototype. The classifier achieved 94.42% accuracy and 99.00% selective accuracy at 89.93% coverage. The segmenter achieved 91.86% deterministic Dice, 92.26% after Monte Carlo averaging, and 93.53% after abstention at 89.36% coverage. The principal contribution is not an isolated high score but a transparent protocol that calibrates predictions, estimates uncertainty, reports rejected cases, and refuses unsupported software inference. The evidence remains internal and image-level; patient-level partitioning, an external cohort, definitive-weight API parity, and clinical reader studies are prerequisites for claims of transportability or clinical validation.

REFERENCES

[1]. Z. U. Abidin, R. A. Naqvi, A. Haider, H. S. Kim, D. Jeong, and S. W. Lee, “Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey,” Frontiers in Bioengineering and Biotechnology, vol. 12, Art. no. 1392807, 2024, doi: 10.3389/fbioe.2024.1392807.

[2]. A. Kurz, K. Hauser, A. Mehrtash, S. Krimmel, H.-P. Schlemmer, K. Maier-Hein, et al., “Uncertainty estimation in medical image classification: Systematic review,” JMIR Medical Informatics, vol. 10, no. 8, Art. no. e36427, 2022, doi: 10.2196/36427.

[3]. M. Musthafa, T. R. Mahesh, V. Vinoth Kumar, and S. Guluwadi, “Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet50,” BMC Medical Imaging, vol. 24, Art. no. 107, 2024, doi: 10.1186/s12880-024-01292-7.

[4]. A. Iqbal, M. A. Jaffar, and R. Jahangir, “Enhancing brain tumour multi-classification using EfficientNet-B0-based intelligent diagnosis for IoMT applications,” Information, vol. 15, no. 8, Art. no. 489, 2024, doi: 10.3390/info15080489.

[5]. B. B. Vimala, S. Srinivasan, S. K. Mathivanan, Mahalakshmi, P. Jayagopal, and G. T. Dalu, “Detection and classification of brain tumor using hybrid deep learning models,” Scientific Reports, vol. 13, Art. no. 23029, 2023, doi: 10.1038/s41598-023-50505-6.

[6]. R. Yousef, S. Khan, G. Gupta, B. M. Albahlal, S. A. Alajlan, and A. Ali, “Bridged-U-Net-ASPP-EVO and deep learning optimization for brain tumor segmentation,” Diagnostics, vol. 13, no. 16, Art. no. 2633, 2023, doi: 10.3390/diagnostics13162633.

[7]. R. Kaifi, “A review of recent advances in brain tumor diagnosis based on AI-based classification,” Diagnostics, vol. 13, no. 18, Art. no. 3007, 2023, doi: 10.3390/diagnostics13183007.

[8]. R. Preetha, J. P. M. Priyadarsini, and J. S. Nisha, “Brain tumor segmentation using multi-scale attention U-Net with EfficientNetB4 encoder for enhanced MRI

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug907

analysis," Scientific Reports, vol. 15, Art. no. 9914, 2025, doi: 10.1038/s41598-025-94267-9.

[9]. S. Iftikhar, N. Anjum, A. B. Siddiqui, M. U. Rehman, and N. Ramzan, "Explainable CNN for brain tumor detection and classification through XAI based key features identification," Brain Informatics, vol. 12, Art. no. 10, 2025, doi: 10.1186/s40708-025-00257-y.

[10]. G. E. S. Shahid et al., "LIU-NET: Lightweight Inception U-Net for efficient brain tumor segmentation from multimodal 3D MRI images," PeerJ Computer Science, vol. 11, Art. no. e2787, 2025, doi: 10.7717/peerj-cs.2787.

[11]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 6105–6114, 2019.

[12]. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.

[13]. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[14]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017, doi: 10.1109/ICCV.2017.74.

[15]. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1321–1330, 2017.

[16]. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proceedings of the 33rd International Conference on Machine Learning, vol. 48, pp. 1050–1059, 2016.

[17]. Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," Advances in Neural Information Processing Systems, vol. 30, 2017.

[18]. M. Nickparvar, "Brain Tumor MRI Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. [Accessed: Aug. 15, 2026].

[19]. A. Akter, "Brain Tumor Segmentation Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/atikaakter11/brain-tumor-segmentation-dataset. [Accessed: Aug. 15, 2026].

[20]. A. S. Tejani, M. E. Klontzas, A. A. Gatti, J. T. Mongan, L. Moy, S. H. Park, et al., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," Radiology: Artificial Intelligence, vol. 6, no. 4, Art. no. e240300, 2024, doi: 10.1148/ryai.240300.

[21]. V. Sounderajah, A. Guni, S. Saria, et al., "The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence," Nature Medicine, 2025, doi: 10.1038/s41591-025-03953-8.

[22]. World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: World Health Organization, 2021, ISBN: 978-92-4-002920-0.

[23]. A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 3868–3878, 2020, doi: 10.1109/TMI.2020.3006437.

[24]. Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," Advances in Neural Information Processing Systems, vol. 32, 2019.

[25]. Y. Ding, J. Liu, X. Xu, M. Huang, J. Zhuang, J. Xiong, et al., "Uncertainty-aware training of neural networks for selective medical image segmentation," Proceedings of Machine Learning Research, vol. 121, pp. 604–615, 2020.

[26]. H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review," npj Digital Medicine, vol. 5, Art. no. 156, 2022, doi: 10.1038/s41746-022-00699-2.

[27]. K. Venkatesh et al., "Gradient-based saliency maps are not trustworthy visual explanations of disease classification in medical imaging," Radiology: Artificial Intelligence, vol. 6, 2024, PMID: 38710971.

[28]. R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, et al., "Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis," npj Digital Medicine, vol. 4, Art. no. 65, 2021, doi: 10.1038/s41746-021-00438-z.

[29]. P. F. Jaeger, C. T. Lüth, L. Klein, and T. J. Bungert, "A call to reflect on evaluation practices for failure detection in image classification," in International Conference on Learning Representations, 2023.

[30]. J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. B. Schön, "Evaluating model calibration in classification," Proceedings of Machine Learning Research, vol. 89, pp. 3459–3467, 2019.

[31]. V. Sounderajah, A. Guni, X. Liu, et al., "Author correction: The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence," Nature Medicine, published Jul. 13, 2026, doi: 10.1038/s41591-026-04570-9.

[32]. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, pp. 199–217, 2021, doi: 10.1038/s42256-021-00307-0.