



IJRTSM

INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

“RELIABILITY-AWARE BRAIN TUMOR CLASSIFICATION FROM MRI: A FRAMEWORK FOR MODEL CALIBRATION, UNCERTAINTY QUANTIFICATION, AND REPRODUCIBLE EVALUATION”

Uma Vishwakarma ^{1*}, Dr. Pritaj Yadav ², Dr. Anurag Jain ³

¹ Research Scholar, Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, India

² Professor, Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, India

³ Professor, Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, India

ABSTRACT

High apparent accuracy on a retrospective brain magnetic resonance imaging (MRI) dataset does not, by itself, demonstrate that a classifier is reliable for clinical use. This paper develops a reliability-oriented extension of modified-Adam convolutional neural network (CNN) brain-tumor detection that complements prior architecture- and optimizer-focused studies. Three accompanying articles and a doctoral thesis were reviewed; their published tabulations were reanalysed without training a new model. A prospective evaluation protocol combines patient-level splitting, source auditing, post-hoc temperature scaling, predictive entropy, selective prediction and domain-shift stress testing. For the binary study, reported model accuracy increased from 91.4% (baseline CNN) and 94.2% (VGG transfer learning) to 97.8% (modified-Adam CNN); these are historical figures rather than a new evaluation. An exploratory four-class confusion matrix contains 759 correct predictions among 783 entries, corresponding to 96.93% matrix-derived accuracy, whereas the same exploratory article lists 97.38% elsewhere. Dataset descriptions also differ between the documents (100-image pilot, 3,000-image binary cohort, approximately 3,100-image multiclass example). The discrepancies illustrate why provenance control and metric recomputation are essential. The proposed reliability protocol specifies calibration error, negative log-likelihood, Brier score, risk-coverage curves and uncertainty-driven referral thresholds. These additional tests are not claimed to have been completed; instead, the study provides a distinct, reproducible evaluation agenda and a transparent audit of existing evidence.

Key Words: Brain MRI; tumor classification; calibration; uncertainty quantification; selective prediction; reproducibility; modified Adam; patient-level validation.

I. INTRODUCTION

Deep learning systems increasingly support interpretation of brain MRI, but their performance must be distinguished from clinical readiness. A classifier can obtain high discrimination on a held-out sample while remaining overconfident, fragile to changes in MRI scanner characteristics, or affected by patient-level leakage. This is especially important when the predictions are intended to guide triage rather than replace qualified interpretation. The earlier Uma Vishwakarma studies addressed CNN architecture, modified Adam optimization and hyperparameter choice in binary and exploratory multiclass MRI classification [1]–[3]. Their central contribution was enhanced predictive performance under the corresponding reported training settings. The present paper asks a different question: what additional evidence is required before a high-scoring model can be regarded as dependable?

Prior foundational work demonstrates that modern neural networks may be miscalibrated even when classification accuracy is high [4]. Under dataset shift, confidence and uncertainty can deteriorate in ways that aggregate accuracy

does not disclose [5]. Transparent prediction-model reporting guidance similarly emphasizes adequate description of data, evaluation and limitations [6]. Consequently, optimizer refinement is necessary but insufficient: trustworthiness also depends on cohort provenance, calibration, selective decision support and external testing.

The objectives are to (i) formulate a leakage-resistant end-to-end reliability evaluation for existing modified-Adam MRI models; (ii) distinguish reported metrics from independently reproducible quantities; (iii) quantitatively audit available tabulations; (iv) specify uncertainty-aware prediction and clinician-referral criteria; and (v) identify the additional evidence required for clinical assessment. Novelty lies in the auditable bridge from optimization performance to reliability assessment—not in an unsupported claim of new classification accuracy.

II. RELATED WORK AND RESEARCH GAP

The first companion article reports a five-convolutional-layer binary network trained with modified Adam on a 3,000-image dataset, partitioned 70:15:15, using 224×224 -pixel inputs, learning rate 0.0005, batch size 32, dropout 0.4 and 50 epochs [1]. It reports 97.8% accuracy and AUC 0.985, accompanied by comparisons and an optimizer ablation. Its optimization results cannot establish calibration, external transportability or patient-level independence without the corresponding prediction files and cohort metadata.

The exploratory multiclass article describes a ResNet50-oriented classifier and an approximate 3,100-image four-class structure with hyperparameter analyses, optimizer comparisons and a four-class confusion matrix [2]. The third article provides a methodological discussion beginning with a 100-image selection and later references the 3,000-image binary dataset [3]. The thesis has related descriptions of the study and methodology [7]. These sources are useful foundations, but they do not describe one fully harmonized cohort or provide the case-level confidence scores required for calibration analyses.

Calibration methods such as temperature scaling learn a single scalar on validation predictions while preserving predicted-class ranking [4]. The study of predictive uncertainty under dataset shift highlights the need for independent distribution-shift evaluation [5]. Reporting standards such as TRIPOD+AI offer a framework for documenting model design, performance and limitations [6]. The unmet need therefore is a reproducible reliability protocol applied with a strict audit trail to MRI classification outputs.

III. MATERIALS AND METHODS

3.1 Study design and evidence classification

The current work is a methodological proposal supported by secondary arithmetic analysis of three provided articles and the associated thesis, rather than a newly executed MRI experiment. “Reported” means copied from an identified earlier source; “recalculated” means deterministically computed from a published table; “proposed” designates a test that requires new predictions or images. No calibration or shift performance is invented.

Table 1. Provenance map derived from the uploaded research documents.

Task / images described	Relevant evidence	Comparability limitation
Binary; 3,000 slices	91.4%, 94.2%, 97.8% model accuracy	Underlying patient IDs and logits absent
Four-class; ~3,100 example	Optimizer table; 4×4 confusion matrix	Exploratory labels; matrix totals 783
100-image pilot; later 3,000-image description	Method and data-preparation narrative	Pilot and main cohort must not be conflated
Doctoral framework	Background and workflow	Cohort identity requires source reconciliation

3.2 Data governance and splitting protocol

For any new prospective evaluation, compile subject identifiers, MRI acquisition site, field strength, pulse sequence, class annotation source, scan date and preprocessing history. Assign all images from an individual patient to exactly one partition before training or augmentation. Duplicate and near-duplicate detection should occur before final splits;

each excluded instance should be logged. Stratify jointly by patient and class whenever possible. Reserve a completely external site or time-separated dataset for model transportability testing. A slice-level 70:15:15 partition alone is not proof of patient independence.

3.3 Reproducible model baseline

Freeze the architecture and checkpoint from the binary modified-Adam CNN when the original weights become accessible. Preserve the input size (224×224), optimizer configuration and originally reported evaluation set as a traceable reference rather than silently merging them with the exploratory ResNet50 experiment. A multiclass model may be separately evaluated, but it requires distinct class mappings and its own held-out results. Store code version, package versions, random seeds, class labels, checkpoint hashes and train/validation/test manifests.

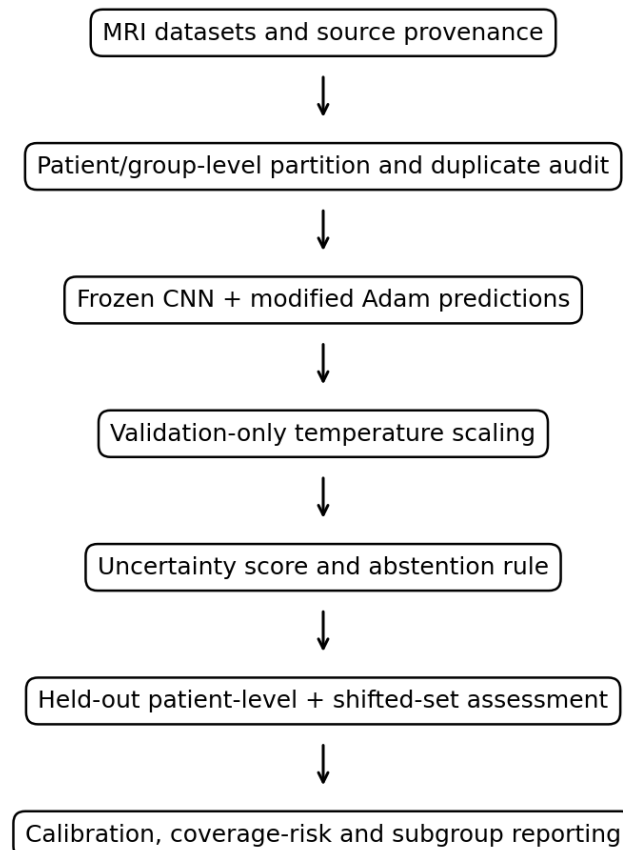


Figure 1. Proposed reliability-assessment pipeline

3.4 Confidence calibration and uncertainty

Given a model logit vector z and temperature $T > 0$, the calibrated probability for class k is $p_k(T) = \exp(z_k/T) / \sum_j \exp(z_j/T)$. Estimate T using validation-set negative log-likelihood only. Never fit T on the held-out test set. Evaluate calibration with reliability diagrams and expected calibration error (ECE), using explicitly stated binning, together with negative log-likelihood (NLL) and multiclass Brier score. The ECE depends on binning; it should not be used alone to make clinical claims. Predictive entropy $H(p) = -\sum_k p_k \ln(p_k)$ serves as one uncertainty proxy, not as a guarantee of true epistemic uncertainty.

3.5 Selective classification

Define confidence $c(x) = \max_k p_k(x)$. An examination is assigned a predicted class only when $c(x) \geq \tau$; otherwise it is referred to a human reviewer. Coverage $C(\tau)$ is the proportion of evaluated scans automatically labelled, and selective risk $R(\tau)$ is the empirical error rate among those accepted. Plot selective risk against coverage across fixed prespecified thresholds; report missed-tumor sensitivity and per-class coverage separately. A system that improves accuracy merely by rejecting difficult cases should also be evaluated for workload, class inequity and potential delay.

3.6 Stress tests and statistical analysis

Prospectively assess perturbation families including moderate Gaussian/Rician noise, bias-field intensity variation, resizing, mild motion blur and scanner/site holdout, ensuring that transformed test cases are never used for hyperparameter tuning. Report accuracy, macro F1, sensitivity, specificity, AUROC where well defined, ECE, NLL, Brier score and selective risk with uncertainty intervals. Use patient-clustered bootstrapping where patient identifiers exist. Where they do not, explicitly limit analysis to image-level descriptive estimates. Pairwise comparisons should use identical cases and report confidence intervals rather than only point estimates.

Table 2. Prespecified reporting metrics for the proposed reliability evaluation.

Measure	Computation / reporting	Interpretation
Accuracy	Correct / total	Overall discrimination; prevalence sensitive
Expected calibration error	$\sum_b (n_b/N) \text{acc}(b) - \text{conf}(b) $	Agreement between confidence and observed success
Brier score	$N^{-1} \sum_i \sum_k (p_{ik} - y_{ik})^2$	Proper probability-quality measure
Negative log-likelihood	$-N^{-1} \sum_i \log p_i y_i$	Penalizes confident errors
Selective coverage	Accepted / all cases	Fraction auto-classified
Selective risk	Errors / accepted cases	Conditional error among automatic decisions
External gap	Metric internal – metric external	Observed transportability difference

IV. RESULTS AND DISCUSSION

4.1 Secondary analysis of reported binary-model results

The binary comparison documented in Paper 1 lists baseline CNN accuracy 91.4%, VGG transfer learning accuracy 94.2%, and proposed modified-Adam CNN accuracy 97.8% [1]. Differences of 6.4 and 3.6 percentage points, respectively, can be computed from those reported values. The same table lists AUC values 0.942, 0.963 and 0.985. These are differences between originally presented configurations, not an independent new trial and not causal proof that the optimizer alone created the full improvement.

Table 3. Source-reported binary performance and arithmetic differences [1].

Model	Reported accuracy (%)	Reported AUC	Difference vs proposed (pp)
Baseline CNN	91.4	0.942	6.4
VGG transfer	94.2	0.963	3.6
Modified-Adam CNN	97.8	0.985	0

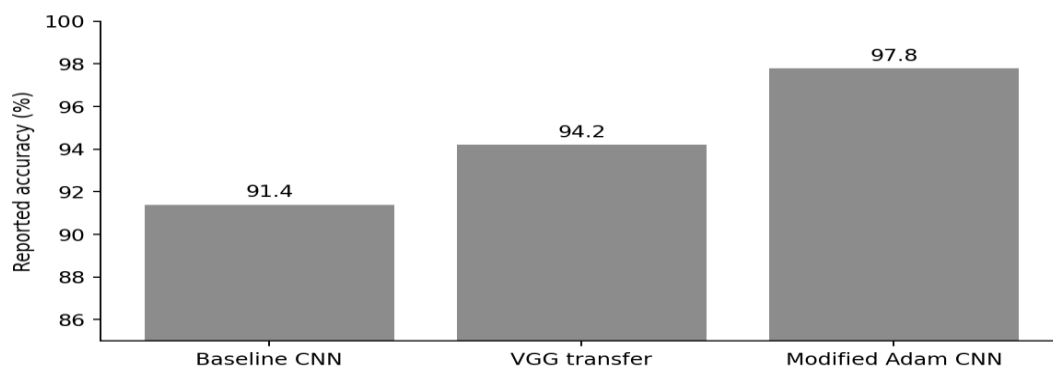


Figure 2. Visualization of previously reported binary-model accuracies [1]; not new test results
<https://www.ijrtsm.com> © International Journal of Recent Technology Science & Management

The binary test split contains 450 slices, and 97.8% is numerically compatible with $440/450 = 97.777\ldots\%$ correct after rounding. However, this calculation cannot uniquely reconstruct a full confusion matrix, precision or recall. The reported accuracy also does not determine the probability calibration or identify which patients, if any, were seen during training.

4.2 Recalculation of the multiclass confusion matrix

The four-class exploratory study presents a matrix whose row totals are 196, 191, 196 and 200 [2]. Summing all 16 entries gives 783. Diagonal entries total $189 + 181 + 192 + 197 = 759$; therefore matrix-derived accuracy is $759/783 \times 100 = 96.9349\%$. The same article lists 97.38% modified-Adam accuracy in a separate optimizer table. The 0.45-percentage-point difference may reflect evaluation of different subsets or editorial inconsistency; the supplied material does not establish which. It must be reconciled before reporting a single coherent multiclass experiment.

Table 4. Reproduced exploratory confusion-matrix counts [2] with recalculated margins.

Actual / predicted	Glioma	Meningioma	Pituitary	No tumor	Row total
Glioma	189	4	2	1	196
Meningioma	5	181	3	2	191
Pituitary	1	2	192	1	196
No tumor	0	1	2	197	200
Column total	195	188	199	201	783

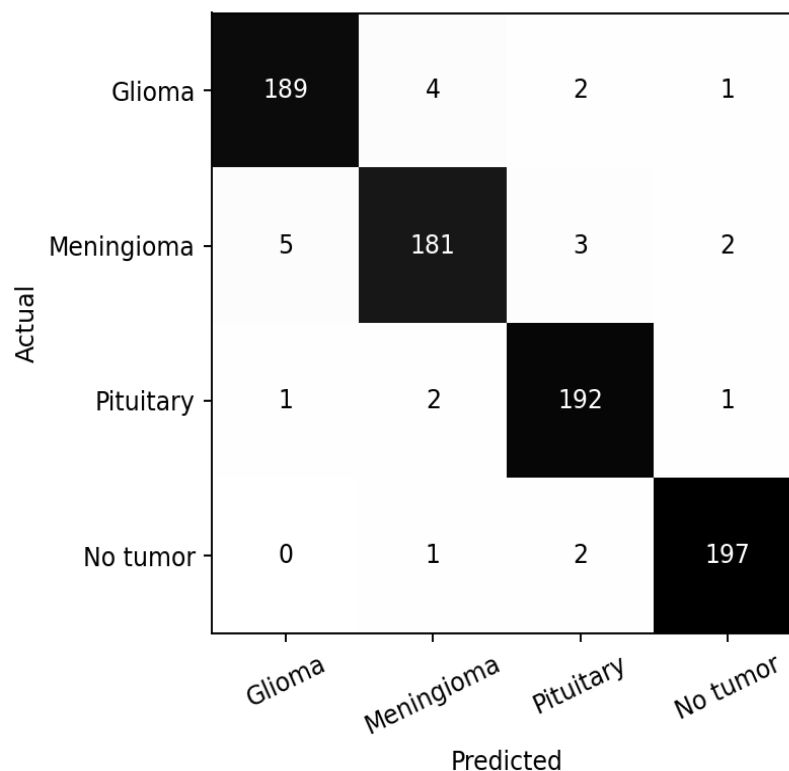


Figure 3. Heatmap reconstructed from the published four-class confusion matrix [2]

From the row totals, class-wise recalls are glioma $189/196 = 96.43\%$, meningioma $181/191 = 94.76\%$, pituitary $192/196 = 97.96\%$, and no tumor $197/200 = 98.50\%$. Their unweighted average is approximately 96.91%. The results are derived from the displayed example matrix and should not be described as independently validated, prospective or externally tested.

Table 5. Metrics deterministically derived from the source multiclass matrix.

Class	Recall from matrix (%)	Support	Incorrect cases
Glioma	96.43	196	7
Meningioma	94.76	191	10
Pituitary	97.96	196	4
No tumor	98.50	200	3
Overall accuracy	96.93	783	24

4.3 Reproducibility discrepancies and their implications

The supporting documents describe a 100-image pilot study, a 3,000-image balanced binary study and an approximately 3,100-image multiclass example. These are not interchangeable. The original dataset name, download version, patient identifiers, random seeds, image-to-patient groupings, calibration predictions and external-cohort manifest are not sufficiently documented in the supplied excerpts to reproduce every claimed test metric independently. In particular, the optimizer comparison and the displayed multiclass matrix do not numerically coincide. These are issues of evidence provenance; they do not prove the original experiments are invalid, but they do prevent a defensible new pooled performance claim.

4.4 Expected outputs of the proposed reliability experiments

If original per-image logits and labels are made available, the first output should be a calibrated-versus-uncalibrated reliability diagram together with NLL, Brier and ECE on untouched tests. A second output should be a risk–coverage curve, with sensitivity and false-negative rates documented at clinically motivated referral thresholds. A third should compare internal with external acquisition conditions. Their numerical results cannot be supplied from the aggregate tables presently available; creating such curves with arbitrary values would be scientifically misleading.

The proposed approach also makes an important distinction between correctly localizing a salient image region and demonstrating clinical explainability. A heatmap is only a visualization of an attribution method; it requires sanity checks, failure-case review and appropriate expert assessment before it can support a claim about clinically relevant lesion boundaries.

V. CONCLUSION

This fourth paper establishes a distinct reliability-focused research direction for the existing modified-Adam MRI tumor classification program. Secondary analysis confirms the earlier binary report of 97.8% accuracy while recognizing it as historical evidence. An independent arithmetic check of the multiclass tabulation gives $759/783 = 96.93\%$ correct and identifies a discrepancy with the separately reported 97.38% metric. The principal contribution is a reproducible blueprint integrating patient-independent splits, calibration, uncertainty-aware referral, distribution-shift assessment and transparent reporting. Before claims of clinical usefulness can be substantiated, the originating case-level files and trained checkpoints must be recovered, dataset inconsistencies resolved, and calibrated external evaluation completed. Thus the manuscript is a defensible methodological and secondary-analysis article—not an unverified new experimental performance study.

V. LIMITATIONS AND FUTURE RESEARCH

The principal limitation is the absence of raw MRI slices, subject IDs, trained-model weights, logits and detailed evaluation records within the four supplied Word documents. Therefore confidence calibration, referral performance, sensitivity under shift, confidence intervals and site-level validation cannot presently be computed. The next experimental phase should obtain the source dataset manifest, reconstruct identical partitions, release transparent training metadata, preregister reliability endpoints and test the fixed model under patient-level and external-domain separation. Medical conclusions should involve qualified radiological oversight and should never rely on this prototype alone.

REFERENCES

- [1] U. Vishwakarma, P. Yadav, and A. Jain, “Deep Learning Based Brain Tumor Detection in MRI Using a Modified Optimizer and Systematic Hyperparameter Tuning,” user-provided manuscript, undated.
- [2] U. Vishwakarma, “Deep Learning Based Brain Tumor Detection with a Modified Optimizer and Fine-Tuned Hyperparameters: An Exploratory Research Article,” user-provided manuscript, 2026.
- [3] U. Vishwakarma, P. Yadav, and A. Jain, “Deep Learning-Based Brain Tumor Detection from MRI Using a Modified Optimizer and Systematic Hyperparameter Tuning,” user-provided manuscript, October 2026.
- [4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR, vol. 70, pp. 1321–1330, 2017. <https://proceedings.mlr.press/v70/guo17a.html>
- [5] Y. Ovadia et al., “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift,” in Advances in Neural Information Processing Systems, vol. 32, 2019. https://papers.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html
- [6] G. S. Collins et al., “TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, 2024, doi: 10.1136/bmj-2023-078378.
- [7] U. Vishwakarma, “Deep Learning based Brain Tumor Detection with a Modified Optimizer and Fine-Tuned Hyper Parameters,” Ph.D. thesis, Rabindranath Tagore University, Bhopal, 2026 (user-provided document).