The Eastasouth Journal of Information System and Computer Science
Vol. 4, No. 01, August 2026, pp. 122-135
ISSN: 3025-566X, DOI: 10.58812/esiscs.v4i01

Reproducibility and Generalization Analysis of OCAT Net S for Oral Cancer Image Classification

Tajul Islam Rafi[1], Md Abu Kawsar Prodhan Hemal[2]
[1,2] Pacific States University

Article Info

Article history:

Received, August, 2026
Revised, August, 2026
Accepted, August, 2026

Keywords:

Calibration;
Clinical Oral Photography;
Deep Learning;
External Validation;
Oral Cancer;
Robustness;
Secondary Experimental Analysis;
Texture-Frequency Gating

ABSTRACT

Clinical photographs could support earlier referral of suspicious oral lesions, but high internal accuracy does not show whether a model will remain reliable when image source, disease prevalence, or lesion definition changes. This study conducted a secondary experimental analysis of the aggregate measurements reported for OCAT-Net-S by [1]. The analysis examined internal discrimination, component ablation, synthetic perturbation tolerance, computational efficiency, probability calibration, threshold behavior, fold-level dependence, leakage controls, and external transfer. No new patient images or model-training runs were introduced. All new results in this paper are derived from the published tables. OCAT-Net-S achieved an internal AUC-ROC of 0.956, sensitivity of 0.942, and oral-cancer F1-score of 0.942 on 1,238 clinical oral images. Its AUC advantage over TinyViT-5M, the strongest evaluated baseline, was 0.015. The external AUC fell to 0.830 on SMART-OM and 0.840 on the Oral Images Dataset, producing absolute generalization gaps of 0.126 and 0.116. These gaps were 8.40 and 7.73 times larger than the internal advantage over TinyViT-5M. Structural ablations caused a mean AUC loss of 0.046, compared with 0.0128 for training-related ablations. Under the most severe Gaussian-noise condition, OCAT-Net-S retained 96.3% of its clean AUC, compared with 93.8% for TinyViT-5M and 90.3% for ResNet-50. The model used 3.55 million parameters and achieved 0.269 AUC per million parameters, although MobileNetV3-Large remained faster and required fewer operations. Calibration was favorable internally, with a Brier score of 0.0419 and ECE-15 of 0.0174. At an assumed 3% screening prevalence, however, estimated positive predictive value ranged from 0.213 to 0.420 across reported operating points. The findings support the internal technical value of texture-frequency spatial gating and compact local-global feature learning. They also show that distribution shift and prevalence exerted a larger practical effect than the margin separating OCAT-Net-S from its strongest internal comparator. Patient-linked multicenter evaluation, fixed external thresholds, prospective workflow testing, and quantitative lesion localization remain necessary before clinical use.

Corresponding Author:

Name: Tajul Islam Rafi
Institution: Pacific States University
Email: p25840@psuca.edu

1. INTRODUCTION

Oral cancer often produces visible changes in mucosal color, texture, contour, ulceration, and lesion boundaries. These changes make clinical photography a practical input for computer-assisted screening, especially in settings where specialist assessment is not immediately available. The same images are difficult to model because camera type, illumination, focus, saliva, dental hardware, anatomical site, and lesion size vary widely. A classifier can perform well by learning disease-related structure, but it can also exploit source-specific cues that do not survive outside the development dataset.

Recent oral-cancer systems combine convolutional layers with attention or transformer components to capture both local surface detail and broader anatomical context [2]. Lightweight designs have become important because screening tools may need to operate on mobile or resource-constrained hardware. Examples include SE-MobileViT, axial-transformer networks, compact convolutional models, and transformer-based diagnostic systems [3]–[5]. Explainable histopathology and segmentation studies have also expanded the field beyond a single image-level score [6]–[8].

[1] introduced OCAT-Net-S, a 3.55-million-parameter model for binary classification of normal and oral-cancer photographs. The network combines convolutional stem, MBConv blocks, squeeze-and-excitation channel recalibration, conditional positional encoding, and Texture Frequency Spatial Gating. The reported evaluation was unusually broad for a compact image classifier. It included grouped cross-validation, repeated seeds, corrected baseline comparisons, ablation tests, calibration, threshold analysis, synthetic corruptions, an image-similarity leakage audit, computational benchmarks, and two external transfer datasets.

The original results contain a second research question beyond model ranking. They allow direct measurement of how architectural gains compare with the larger losses caused by image corruption, prevalence change, and external distribution shift. That comparison matters because a small internal improvement can appear statistically convincing while remaining minor relative to the uncertainty introduced when the model is moved to another setting. A focused reanalysis can also separate parameter efficiency from arithmetic speed, distinguish probability calibration from clinical usefulness, and quantify the dependence created by repeated training on the same folds.

This study therefore reanalyzed the published OCAT-Net-S measurements as a unified experimental record. The objectives were to quantify the internal margin over reproduced baselines, compare structural and training ablations, calculate AUC retention under perturbation, examine efficiency tradeoffs, evaluate threshold behavior at lower prevalence, measure fold-level dependence, and contrast internal performance with external transfer. The study did not reproduce training or claim access to patient-level records. Its contribution is a set of derived comparisons that place the reported findings on a common practical scale.

Although convolutional neural networks effectively capture local visual patterns, their ability to model long-range relationships across lesion regions remains limited. Recent oral-cancer classification studies have therefore adopted transformer-based architectures to capture both localized lesion characteristics and broader spatial context [9]. For example, [10] introduced a hierarchical lesion-aware transformer specifically designed for oral-cancer image classification. Nevertheless, further research is needed to develop computationally efficient models that integrate texture, frequency, and spatial information while maintaining reliable classification performance.

2. MATERIALS AND METHODS

2.1 Study Design

We performed a secondary experimental analysis of the methods and aggregate results reported by [1]. The source article was treated as an

experimental record. Numerical values were extracted from its dataset, baseline, ablation, robustness, efficiency, calibration, threshold, dependence, leakage-audit, and external-transfer tables. Values were entered into structured analysis matrices and checked against the surrounding results text. The unit of analysis was a reported model, ablation condition, perturbation condition, fold, operating point, or evaluation cohort.

This design differs from a literature review because it tests prespecified quantitative questions using one complete experimental record. It also differs from a primary replication because the trained weights, raw logits, patient identifiers, and complete image-level predictions were not re-created. Inferential values such as adjusted p-values and confidence intervals were retained from the source. New calculations were descriptive transformations of published aggregate results.

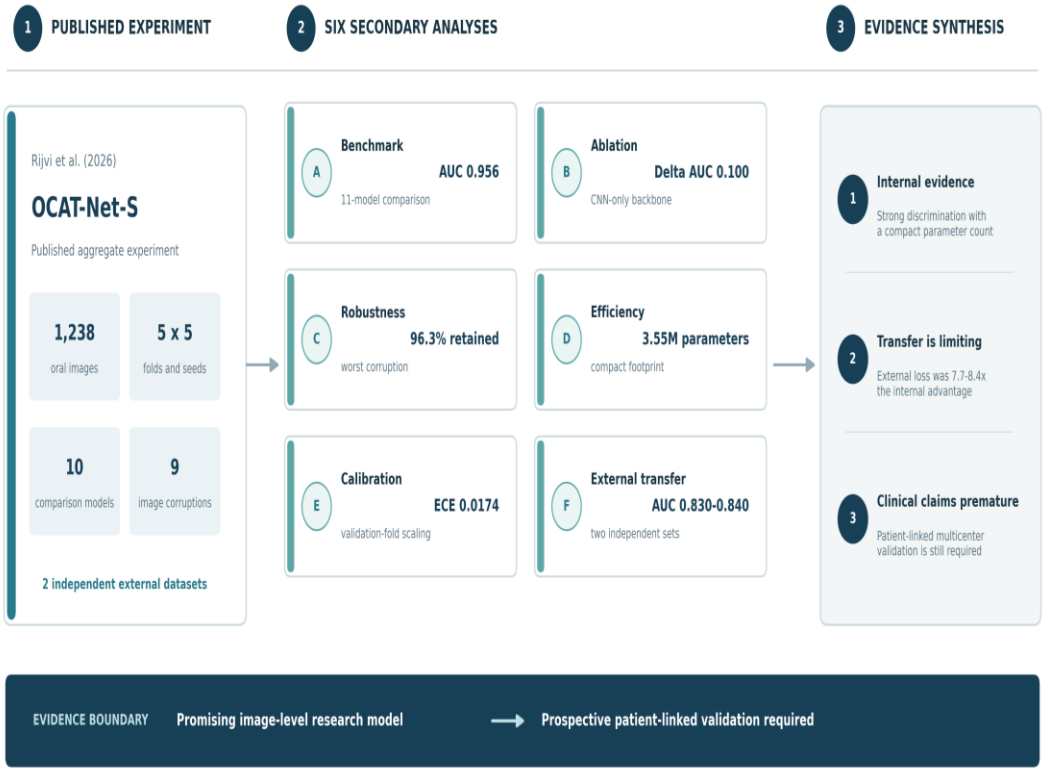

Figure 1. Evidence pathway for the secondary experimental analysis of OCAT-Net-S. Published aggregate measurements were organized into six analytic domains. The numeric summaries show internal performance, component effects, robustness, efficiency, calibration, and external transfer before the evidence boundary was defined.

2.2 *Source Dataset and Classification Task*

The internal reference dataset contained 1,238 RGB clinical oral-cavity images from the Atef-Kaggle Oral Cancer Images for Classification collection. The binary labels included 553 normal images and 685 oral-cancer images, corresponding to 44.7% and 55.3% of the dataset. Patient identifiers were unavailable. [1] used five-fold stratified group cross-validation with filename-sequential grouping as a proxy for clinical sessions. Five training seeds produced 25-fold-seed evaluations. This design reduced the chance that visually related files would cross folds, but it did not establish patient-level independence.

Two independent datasets were used for transfer testing without external threshold tuning. SMART-OM contained

2,145 normal images and 20 oral squamous cell carcinoma images, giving a positive prevalence of 0.9%. The Oral Images Dataset contained 165 benign and 158 malignant images, giving a positive prevalence of 48.9%. The external tasks differed from the internal task in source, class definition, prevalence, and negative-class composition. External results were therefore analyzed as transfer measurements rather than exact replications.

Table 1. Experimental Data Sources and Analysis Roles

Source	Sample or unit	Primary role	Key limitation
Atef Kaggle internal dataset	1,238 images	Grouped internal cross-validation	No patient identifiers
Baseline matrix	11 models	Internal model comparison	Limited to reproduced comparators
Ablation matrix	Full model and 8 variants	Component contribution	Single-component removals are not additive
Perturbation matrix	3 models across 9 corruptions	Controlled tolerance analysis	Synthetic corruption is not clinical shift
SMART OM	2,165 images	External transfer	Only 20 positive cases
Oral Images Dataset	323 images	External transfer	Different benign versus malignant task

2.3 OCAT Net S Architecture

OCAT-Net-S uses a hierarchical sequence of local and global feature operations. A convolutional stem performs early downsampling and local edge extraction. MBConv blocks expand and project feature channels with depthwise convolution. MBConv-SE blocks add squeeze-and-excitation weights so the network can recalibrate channels before later processing. Conditional positional encoding supplies spatial position information without a fixed learned position table. The Texture Frequency Spatial Gating module operates in later stages and combines texture-sensitive frequency information with spatial gating. The classifier converts the final representation into a probability for the oral-cancer class.

The source model contained 3.55 million parameters and required 1.478 GFLOPs per image. Training used AdamW, layer-wise learning-rate decay, gradient clipping, stochastic depth, focal loss, label smoothing, MixUp, CutMix, and an exponential moving average of the weights. Temperature scaling was fitted on validation-fold logits and applied unchanged to the held-out fold. These choices formed the full reference condition for the ablation analysis.

Table 2. Reconstructed OCAT Net S Experimental Protocol

Element	Reported specification	Purpose in the experiment
Task	Normal versus oral cancer	Binary image-level classification
Validation	Five grouped folds and five seeds	Estimate internal discrimination and training variability
Grouping	Filename-sequential proxy groups	Reduce visually related cross-fold leakage
Optimizer	AdamW with weight decay 0.05	Stable fine-tuning and regularization
Learning rates	Layer-wise decay factor 0.75	Smaller updates in earlier stages
Loss	Focal loss with label smoothing	Address difficult examples and confidence
Augmentation	MixUp and CutMix	Reduce overfitting
Calibration	Validation-only temperature scaling	Improve probability interpretation
External testing	Fixed internal thresholds	Measure transfer without target tuning

2.4 Outcomes and Derived Measures

The primary performance outcome was AUC-ROC. Secondary outcomes were positive-class F1-score, sensitivity, specificity, balanced accuracy, predictive values, likelihood ratios, calibration error, negative log likelihood, parameter count, GFLOPs, latency, throughput, memory use, and disk size. Reported confidence intervals, corrected p-values, and Cohen d values were used to judge the baseline and ablation contrasts.

We calculated an absolute AUC gap by subtracting the comparison score from the reference score. AUC retention expressed a perturbed or external score as a proportion of the clean internal score. Relative degradation was one minus retention. Parameter efficiency was read from the source and checked as AUC divided by millions of parameters. We also calculated the ratio between each external generalization gap and the internal advantage over TinyViT-5M. This ratio shows whether the gain that separated the proposed model from its strongest comparator was large or small relative to dataset shift.

$$\text{Absolute gap} = \text{AUC reference} - \text{AUC comparison}$$

$$\text{AUC retention} = \text{AUC shifted} / \text{AUC clean}$$

$$\text{Generalization to margin ratio} = \text{external AUC gap} / 0.015$$

Ablations were assigned to two descriptive groups. Structural ablations changed the backbone, TFSG, MBConv-SE, or positional encoding. Training ablations changed augmentation, loss, label smoothing, or the exponential moving average. We compared the mean AUC loss in these groups. Because each ablation was run separately, the mean losses were not summed and were not interpreted as independent causal shares.

For the leakage audit, we calculated within-to-cross-group ratios for the proportion of pairs meeting the source thresholds for perceptual hash, structural similarity, and CLIP cosine similarity. For repeated-fold dependence, we retained the reported variance components and fold intraclass correlation. Threshold results were compared at an assumed screening prevalence of 3%, with attention to positive predictive value and false negatives per 1,000 screened. All calculations used unrounded source values when they were available; otherwise, the reported three-decimal values were used.

2.5 Statistical Interpretations

The source study used paired comparisons within a shared cross-validation protocol and applied Holm-Bonferroni correction. The present analysis did not recompute those tests because fold-level predictions were unavailable. A contrast was treated as statistically supported only when the source reported an adjusted p-value below .05 and a confidence interval that excluded zero. Descriptive ratios were used to compare magnitudes across experiments. They were not assigned new p-values because the underlying observations were neither independent nor available at image level.

Repeated seeds on the same folds were treated as training replications, not as new clinical samples. External cohorts were interpreted separately because their class definitions and prevalence differed. No causal claim was made about clinical outcomes, and no estimate was treated as a diagnostic accuracy measure for a target population unless the source design supported that interpretation.

3. RESULTS AND DISCUSSION

3.1 Internal Discrimination

OCAT-Net-S achieved the highest internal AUC-ROC and oral-cancer F1-score among the 11 evaluated models. Its AUC-ROC was 0.956, with a reported standard deviation of 0.014, and its oral-cancer F1-score was 0.942.

TinyViT-5M was the strongest AUC comparator at 0.941, producing an absolute margin of 0.015. EfficientViT-M4 had the next highest AUC at 0.935 and the strongest baseline F1-score at 0.901.

The advantage over TinyViT-5M was statistically supported in the source analysis, with a 95% confidence interval of 0.006 to 0.024, adjusted p = .0013, and Cohen d = 1.00. The largest AUC difference was 0.058 against ResNet-50. The ranking remained favorable across the listed comparators, but the practical margin narrowed as the baseline architecture became stronger.

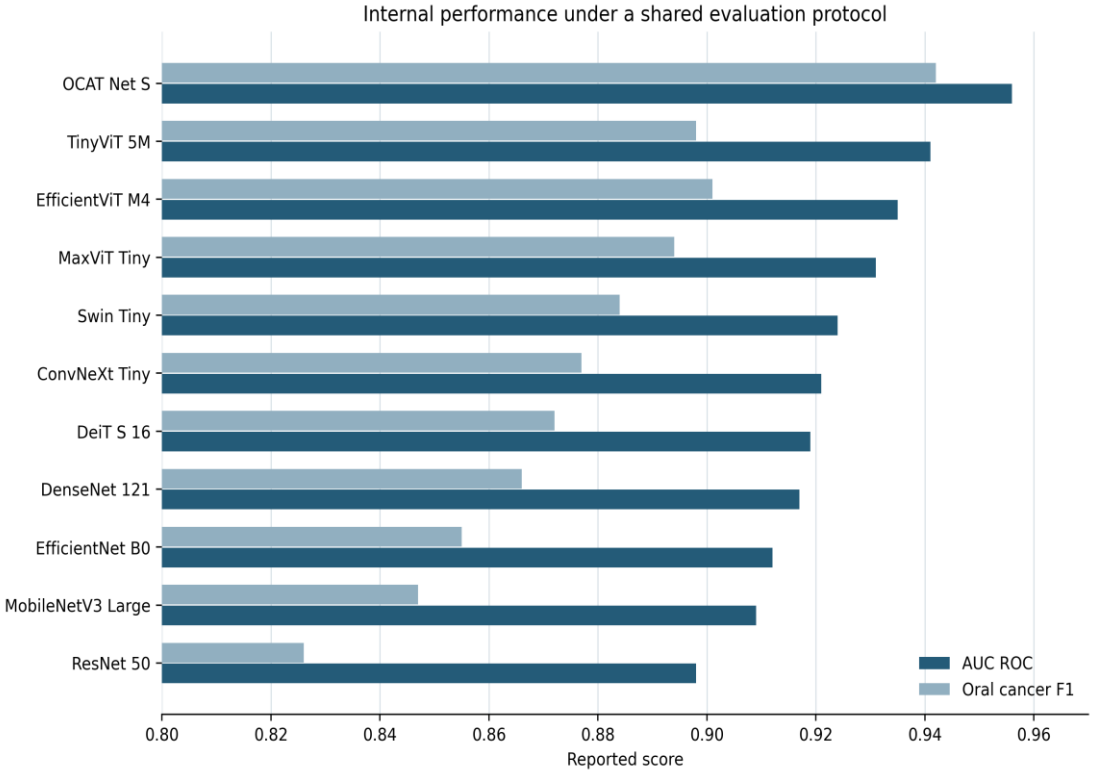

Figure 2. AUC-ROC and oral-cancer F1-scores reported under the same grouped cross-validation protocol. Values originate from [1].

Table 3. Internal baseline comparison

Model	AUC ROC	Oral cancer F1	Delta AUC versus OCAT Net S	Parameters M
OCAT Net S	0.956	0.942	Reference	3.55
TinyViT 5M	0.941	0.898	0.015	5.40
EfficientViT M4	0.935	0.901	0.021	8.80
MaxViT Tiny	0.931	0.894	0.025	30.90
Swin Tiny	0.924	0.884	0.032	28.30
ConvNeXt Tiny	0.921	0.877	0.035	28.60
DeiT S 16	0.919	0.872	0.037	22.10
DenseNet 121	0.917	0.866	0.039	8.00
EfficientNet B0	0.912	0.855	0.044	5.30
MobileNetV3 Large	0.909	0.847	0.047	5.40
ResNet 50	0.898	0.826	0.058	25.60

3.2 *Component Contribution*

The ablation pattern separated architectural effects from smaller training refinements. Reducing the network to a CNN-only backbone lowered AUC-ROC from 0.956 to 0.856, a loss of 0.100. Replacing TFSG with standard multi-head self-attention lowered AUC-ROC by 0.043. Replacing MBConv-SE with MBConv caused a loss of 0.027. These three contrasts had adjusted p-values below .001 and Cohen d values from 1.54 to 4.52.

Removing MixUp and CutMix reduced AUC-ROC by 0.021. The losses associated with conditional positional encoding, focal loss, and label smoothing ranged from 0.012 to 0.014. Removing the exponential moving average reduced AUC-ROC by 0.005, and this difference was not statistically significant after correction. The mean loss across four structural changes was 0.0460. The mean across four training changes was 0.0128. The structural mean was therefore 3.61 times the training mean. This comparison is descriptive because the ablations were not factorial and may interact.

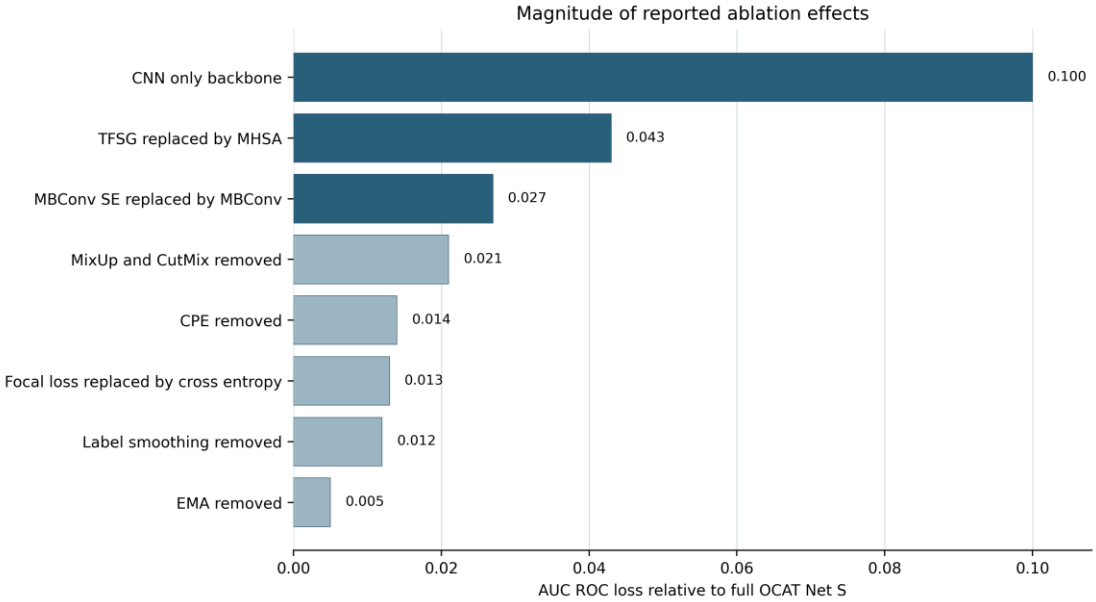

Figure 3. AUC-ROC loss for each reported ablation relative to the full model. Larger bars indicate a larger reduction after the component was changed or removed.

Table 4. Reanalysis of Component Ablations

Ablation	Variant AUC	AUC loss	Delta MCC	Adjusted p	Cohen d
CNN only backbone	0.856	0.100	0.198	< .001	4.52
TFSG replaced by MHSA	0.913	0.043	0.094	< .001	2.65
MBConv SE replaced by MBConv	0.929	0.027	0.058	< .001	1.54
MixUp and CutMix removed	0.935	0.021	0.044	< .001	1.25
CPE removed	0.942	0.014	0.031	.0144	0.88
Focal loss replaced by cross entropy	0.943	0.013	0.028	.0195	0.83
Label smoothing removed	0.944	0.012	0.026	.0264	0.75
EMA removed	0.951	0.005	0.009	.180	0.34

3.3 *Synthetic Perturbation Tolerance*

All three models lost performance as corruption severity increased. OCAT-Net-S retained the highest AUC-ROC in every reported condition. Its lowest AUC was 0.921

under high Gaussian noise. Relative to its clean AUC of 0.956, worst-case retention was 96.3% and degradation was 3.7%. TinyViT-5M retained 93.8% of its clean AUC in its worst condition, while ResNet-50 retained 90.3%.

The separation between OCAT-Net-S and TinyViT-5M increased as Gaussian noise became stronger. The clean AUC gap was 0.015. It rose to 0.017 at low noise, 0.027 at medium noise, and 0.038 at high noise when calculated from the displayed point estimates. Under high blur and high JPEG compression, the corresponding gaps were 0.022 and 0.018. These results indicate tolerance to the tested corruptions, although synthetic noise, blur, and compression cannot represent all changes encountered in clinical photography.

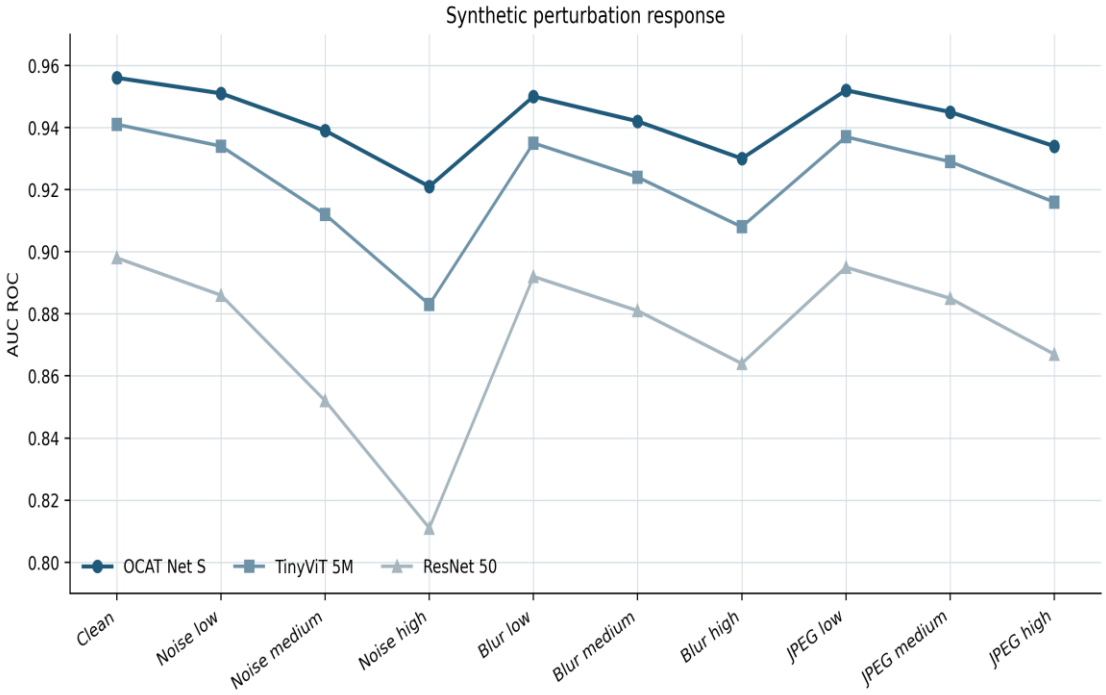

Figure 4. AUC-ROC under clean input and nine synthetic perturbation conditions. The sequence groups Gaussian noise, Gaussian blur, and JPEG compression by severity.

Table 5. Worst Case Perturbation Summary

Model	Clean AUC	Worst AUC	Worst condition	Retention	Relative degradation
OCAT Net S	0.956	0.921	Gaussian noise high	96.3%	3.7%
TinyViT 5M	0.941	0.883	Gaussian noise high	93.8%	6.2%
ResNet 50	0.898	0.811	Gaussian noise high	90.3%	9.7%

3.4 Computational Efficiency

OCAT-Net-S contained fewer parameters than every reproduced comparator and had the smallest reported disk size and VRAM requirement. Its AUC per million parameters was 0.269, compared with 0.174 for TinyViT-5M. This represents a 54.6% higher parameter-efficiency ratio while using 34.3% fewer parameters. OCAT-Net-S also produced 625 images per second in the reported batch benchmark, compared with 556 for TinyViT-5M.

Parameter efficiency did not make OCAT-Net-S the fastest model. MobileNetV3-Large required 0.22 GFLOPs, processed 1,111 images per

second, and achieved 4.132 AUC per GFLOP. EfficientNet-B0 and EfficientViT-M4 also used fewer operations and had higher throughput. The proposed model therefore occupied a favorable region for accuracy and model size, while simpler mobile architecture remained preferable when arithmetic cost or latency dominated the design constraint.

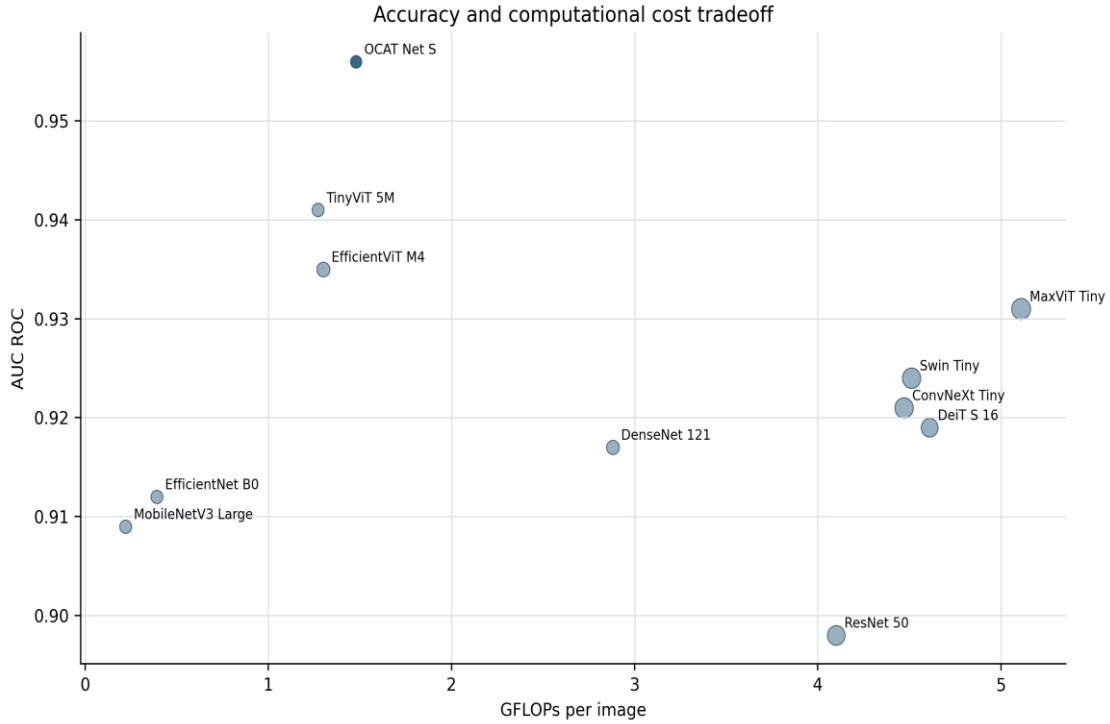

Figure 5. Relationship between computational cost and internal AUC-ROC. Marker area is proportional to parameter count. OCAT-Net-S combines high AUC with a small parameter footprint, but it is not the lowest-GFLOP model.

Table 6. Selected Efficiency Comparisons

Model	Parameters M	GFLOPs	Single image ms	Throughput images s	AUC per M parameter	AUC per GFLOP
OCAT Net S	3.55	1.478	6.4	625	0.269	0.647
TinyViT 5M	5.40	1.27	7.1	556	0.174	0.741
EfficientViT M4	8.80	1.30	4.8	909	0.106	0.719
MobileNetV3 Large	5.40	0.22	4.2	1,111	0.168	4.132
EfficientNet B0	5.30	0.39	5.6	833	0.172	2.338
ResNet 50	25.60	4.10	8.2	500	0.035	0.219

3.5 Calibration and Threshold Behavior

Validation-only temperature scaling produced a mean Brier score of 0.0419, ECE-15 of 0.0174, and negative log likelihood of 0.1798. The mean fitted temperature was 1.34. At the F1-optimal threshold of 0.418, sensitivity was 0.942 and specificity was 0.928. The calibration results support internally coherent probability estimates, but calibration on development folds does not guarantee calibration in a new population.

Prevalence had a pronounced effect on predictive value. At an assumed screening prevalence of 3%, the high-sensitivity operating point yielded a

positive predictive value of 0.213 and approximately 1.20 false negatives per 1,000 people screened. The F1-optimal point yielded a positive predictive value of 0.288 and 1.74 false negatives per 1,000. The high specificity point increased positive predictive value to 0.420 but produced 3.30 false negatives per 1,000. Negative predictive value remained above 0.996 across these scenarios. These calculations show why a threshold selected on a dataset with 55.3% positive images cannot be adopted directly for population screening.

Table 7. Threshold Behavior at Assumed 3 Percent Prevalence

Operating point	Threshold	Sensitivity	Specificity	PPV	NPV	False negatives per 1000
High sensitivity	0.310	0.960	0.890	0.213	0.9986	1.20
Youden optimal	0.460	0.930	0.945	0.343	0.9977	2.10
F1 optimal	0.418	0.942	0.928	0.288	0.9981	1.74
High specificity	0.550	0.890	0.962	0.420	0.9965	3.30

3.6 Dependence and Leakage Audit

Fold means ranged from 0.942 to 0.972, while the grand mean was 0.956. The source variance decomposition attributed 70.4% of component variance to folds, 27.9% to seeds, and 1.6% to fold-by-seed residual variation. The fold intraclass correlation was 0.704. Most of the observed variation therefore followed the data split rather than stochastic retraining. Treating all 25 fold-seed values as independent observations would overstate the effective amount of evidence.

The image-similarity audit showed that related images were concentrated within the filename-based proxy groups. The near-duplicate rate at perceptual-hash distance of five or less was 9.3% within groups and 0.4% across groups, a ratio of 23.25. The rate at structural similarity of at least 0.90 was 11.8% within groups and 0.5% across groups, a ratio of 23.6. CLIP cosine similarity of at least 0.95 occurred in 8.1% of within-group pairs and 0.2% of cross-group pairs, a ratio of 40.5. The grouping strategy appears to have reduced visual overlap, but image similarity cannot establish that images came from different patients.

Table 8. Dependence and Leakage Indicators

Indicator	Within or fold value	Cross group or comparison	Interpretation
Fold variance component	70.4%	Seed component 27.9%	Data split dominated variability
Fold ICC	0.704	Residual component 1.6%	Repeated seeds were clustered within folds
pHash distance 5 or less	9.3%	0.4% across groups	23.25 times higher within groups
SSIM 0.90 or greater	11.8%	0.5% across groups	23.6 times higher within groups
CLIP similarity 0.95 or greater	8.1%	0.2% across groups	40.5 times higher within groups

3.7 External Transfers

External transfer produced the clearest performance change. AUC-ROC declined from 0.956 internally to 0.830 on SMART-OM and 0.840 on the Oral Images Dataset. The absolute gaps were

0.126 and 0.116, corresponding to AUC retention of 86.8% and 87.9%. The SMART-OM gap was 8.40 times the internal advantage over TinyViT-5M. The Oral Images gap was 7.73 times that advantage.

The effect was larger for threshold-dependent outcomes. On SMART-OM, sensitivity was 0.660 and positive-class F1 was 0.163, although specificity remained 0.940 and negative predictive value was 0.997. The extremely low positive prevalence and the presence of only 20 positive cases limited precision. On the Oral Images Dataset, sensitivity was 0.790, specificity was 0.800, and positive-class F1 was 0.790. This more balanced result remained below internal performance and addressed a different benign-versus-malignant task.

The likelihood ratios clarify the same pattern. SMART-OM produced a positive likelihood ratio of 11.00 and a negative likelihood ratio of 0.36. The model could therefore increase suspicion after a positive result, but its sensitivity was not high enough to exclude disease after every negative result. The Oral Images Dataset produced a lower positive likelihood ratio of 3.95 and a negative likelihood ratio of 0.26. These values describe cohort-specific evidence and should not be applied to a screening population with different case selection.

The external findings also reveal why AUC and fixed-threshold metrics should be reported together. AUC measures ranking across possible thresholds, whereas F1-score, predictive values, and referral burden depend on one operating point and the cohort prevalence. The SMART-OM AUC retained 86.8% of the internal value, yet the positive-class F1-score retained only 17.3% of its internal value. This divergence shows that partial ranking ability can coexist with poor operational performance when the threshold and prevalence move away from the development setting.

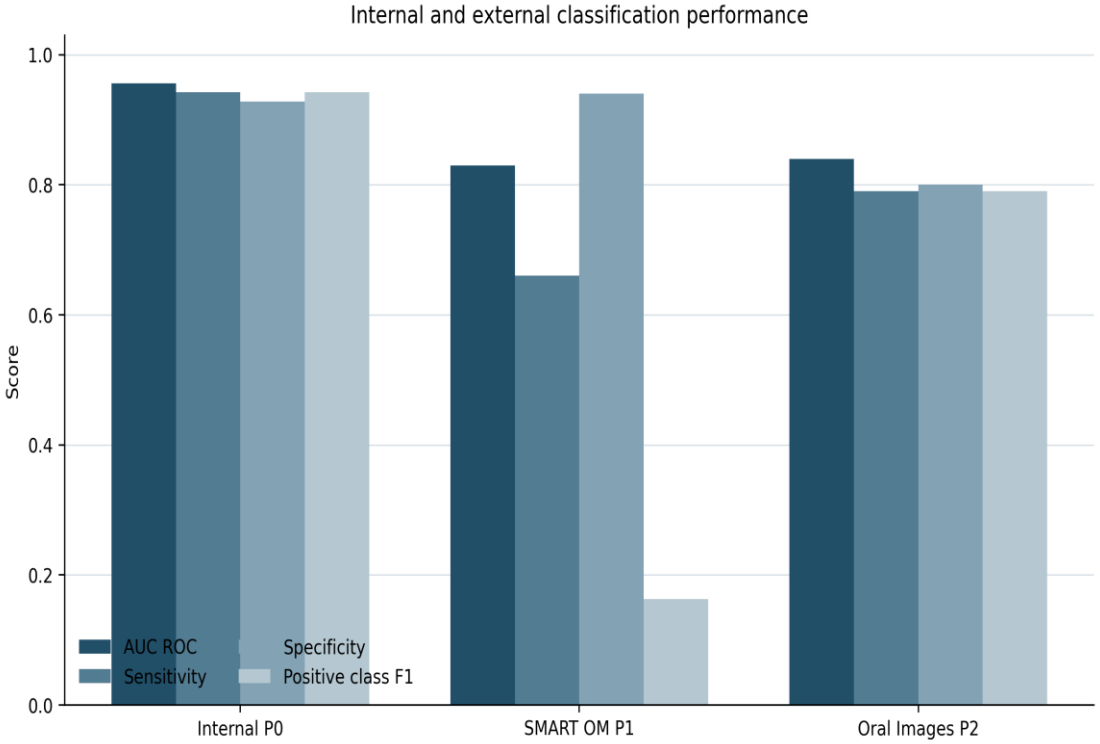

Figure 6. Internal and external performance using thresholds selected from internal validation folds. Differences in prevalence and class definition require cohort-specific interpretation.

Table 9. Internal and External Performance

Cohort	Positive prevalence	AUC ROC	Sensitivity	Specificity	PPV	NPV	Positive F1
Internal P0	0.553	0.956	0.942	0.928	0.942	0.928	0.942
SMART OM P1	0.009	0.830	0.660	0.940	0.093	0.997	0.163
Oral Images P2	0.489	0.840	0.790	0.800	0.791	0.799	0.790

4. DISCUSSION

4.1 Main Findings

The reanalysis supports three main conclusions. First, OCAT-Net-S achieved the strongest internal discrimination among the reproduced models while using the fewest parameters. Second, the ablation results linked most of the internal gain to the hybrid architecture, TFSG, and channel recalibration rather than to any single training refinement. Third, the loss under external transfer was much larger than the gain over the strongest internal baseline. The third result is the most important for interpretation because it changes the practical meaning of a 0.015 internal AUC advantage.

The ablation evidence is consistent with the visual structure of oral lesions. Local texture and boundaries matter, but a lesion also needs to be interpreted within surrounding mucosa and anatomical context. A compact network that combines convolutional texture extraction with gated spatial processing can represent both scales. The 0.043 AUC loss after replacing TFSG with standard self-attention suggests that the frequency-spatial interaction added useful information within this architecture. It does not prove that TFSG will outperform attention in every dataset or implementation.

Related hybrid systems in medical imaging offer a plausible cross-domain context. [11] combined a CNN with a Swin Transformer for explainable tuberculosis screening, while [12] fused CoAtNet and frozen DINOv2 features through channel attention for chest radiographs. These studies support the broader use of selective local-global fusion, but their numerical results cannot be transferred to oral photography because anatomy, acquisition, labels, and clinical pathways differ.

4.2 Internal Performance and External Validity

The internal confidence interval and corrected comparisons show that OCAT-Net-S was not merely the highest point estimate in the reported benchmark. However, internal significance answers a narrow question: whether the models differed under the same folds, seeds, and image source. External validity asks whether the learned relationship holds when those conditions change. The 0.116 to 0.126 external AUC losses were nearly eight times the internal margin over TinyViT-5M, showing that data source affected performance more strongly than the choice between the two leading internal models.

The external results should not be reduced to one claim of failure. An AUC of 0.830 or 0.840 still indicates partial ranking ability, and the SMART-OM result retained high specificity. The main concern is threshold portability. With only 20 positive SMART-OM cases, positive predictive value fell to 0.093 and F1-score to 0.163. A model used for referral could therefore generate many false alerts relative to true positives, even when specificity appears high. External recalibration might improve operation, but tuning on the target set would convert a transfer test into a new development exercise.

4.3 Robustness and Efficiency

The perturbation experiment adds useful evidence because OCAT-Net-S degraded more slowly than TinyViT-5M and ResNet-50 under severe

Gaussian noise. The model also retained a smaller but consistent advantage under blur and JPEG compression. Frequency-guided gating may help preserve discriminative structure when fine image detail is damaged. This interpretation remains provisional because the source did not isolate a frequency mechanism under each perturbation and because synthetic corruption is simpler than real camera and workflow variability.

Efficiency should be described with more than parameter count. OCAT-Net-S stored a strong representation in 3.55 million parameters and achieved the highest AUC per million parameters. MobileNetV3-Large used far fewer operations and ran faster. A mobile screening system may prefer the faster network if battery use and latency are binding. A server-assisted workflow may prefer OCAT-Net-S if the higher internal AUC and perturbation tolerance remain stable after external validation. The correct choice depends on measured constraints in the intended workflow.

4.4 Calibration and Screening Use

The internal calibration results were strong, particularly the ECE-15 of 0.0174. Calibration makes a predicted probability easier to interpret within the population used to fit and evaluate the scaling function. It does not remove prevalence effects. The threshold analysis demonstrates this distinction. At 3% prevalence, a high-sensitivity setting had an estimated positive predictive value of 0.213. Most positive alerts would still be false positives, although the negative predictive value would remain high.

Screening thresholds should therefore be selected with the intended referral capacity, disease prevalence, and consequences of missed lesions in view. A high-sensitivity threshold reduced estimated false negatives but increased referrals. A high-specificity threshold improved positive predictive value but nearly tripled false negatives per 1,000 compared with the high-sensitivity setting. Prospective evaluation should report the number of referrals, confirmed disease cases, missed cancers, time to specialist assessment, and subgroup performance rather than relying on AUC alone.

4.5 Validity Limits

This analysis inherits the limitations of the source experiment. The internal dataset lacked patient identifiers, biopsy-confirmation metadata, lesion masks, and complete acquisition information. Filename-based grouping reduced measured visual similarity across folds but cannot prove patient separation. Grad-CAM images can show where the network responds, yet they do not validate lesion localization or clinical reasoning. The external cohorts used different tasks and prevalence, which is useful for stress testing but prevents a direct estimate of generalization to one defined clinical population.

The present study also has limits. It relied on aggregate tables rather than raw predictions and did not rerun model training. Derived ratios are affected by rounding. Structural and training ablation means describe magnitude but do not estimate independent causal contributions. The external-to-internal comparisons combine distribution shift with changes in class definition and prevalence. These constraints are why the paper reports a secondary experimental analysis rather than a primary replication.

4.6 Requirements For A Confirmatory Study

A confirmatory study should enroll patients prospectively at several clinical sites and assign stable patient and lesion identifiers. Training, validation, and test sets should be separated by patient and preferably by site. The reference standard should record histopathology when available and document how normal, benign, potentially malignant, and malignant classes were assigned. Acquisition metadata should include device,

illumination, distance, focus quality, anatomical site, and operator [13].

The model should be frozen before external evaluation. Calibration and referral thresholds should be prespecified or updated under a documented protocol. The analysis should report confidence intervals, decision curves, calibration plots, subgroup results, and clinician-reviewed errors. Lesion masks or bounding boxes would permit quantitative localization measures. A final prospective phase should compare usual practice with model-assisted assessment and record referral yield, missed disease, review time, and clinician override behavior.

5. CONCLUSION

OCAT-Net-S combined high internal discrimination with a small parameter footprint. Its strongest reported advantages came from the hybrid backbone, texture-frequency spatial gating, and channel recalibration. The model also retained more of its clean AUC than two comparator networks under severe synthetic corruption. These findings justify further testing of architecture.

The external results define the current boundary of the evidence. AUC fell by 0.116 to 0.126 across two independent datasets, a loss far larger than the 0.015 advantage over the strongest internal baseline. Calibration and threshold analyses further showed that low prevalence can sharply reduce positive predictive value. OCAT-Net-S should therefore be treated as a promising image-level research model, not as a validated diagnostic system. Patient-linked multicenter data and prospective workflow evaluation are required before clinical claims can be made.

REFERENCES

[1] M. H. Rijvi *et al.*, "A compact texture frequency spatial gating network for oral cancer classification from clinical oral images," *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-59480-0.

[2] A. L. D. Araújo, C. M. Pedroso, P. A. Vargas, M. A. Lopes, and A. R. Santos-Silva, "Advancing oral cancer diagnosis and risk assessment with artificial intelligence: a review," *Explor. Digit. Heal. Technol.*, vol. 3, p. 101147, 2025, doi: 10.37349/edht.2025.101147.

[3] M. F. Kabir, M. Y. Ahmad, R. Uddin, M. Cordero, and S. Kant, "Accurate and lightweight oral cancer detection using SE-MobileViT on clinically validated image dataset," *Discov. Artif. Intell.*, vol. 5, p. 173, 2025, doi: 10.1007/s44163-025-00442-2.

[4] D. P. Yadav, B. Sharma, A. Noonia, and A. Mehbodniya, "Explainable label guided lightweight network with axial transformer encoder for early detection of oral cancer," *Sci. Rep.*, vol. 15, no. 1, p. 6391, 2025, doi: 10.1038/s41598-025-87627-y.

[5] P. Liu and K. Bagi, "A tailored deep learning approach for early detection of oral cancer using a 19-layer CNN on clinical lip and tongue images," *Sci. Rep.*, vol. 15, no. 1, p. 23851, 2025, doi: 10.1038/s41598-025-07957-9.

[6] F. R. D. Cruze, J. Wasima, M. F. Hosen, M. B. A. Miah, Z. Muhammad, and M. F. Al Masud, "Oral Cancer Diagnosis Using Histopathology Images: An Explainable Hybrid Transformer Framework," *Technologies*, vol. 14, no. 1, p. 39, 2026, doi: 10.3390/technologies14010039.

[7] A. Hossain *et al.*, "Transformer-Based Ensemble Model for Binary and Multiclass Oral Cancer Segmentation," *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2025. doi: 10.1109/ECCE64574.2025.11012921.

[8] J. Štifanić, D. Štifanić, N. Anđelić, and Z. Car, "Explainable AI for Oral Cancer Diagnosis: Multiclass Classification of Histopathology Images and Grad-CAM Visualization," *Biology (Basel).*, vol. 14, no. 8, p. 909, 2025, doi: 10.3390/biology14080909.

[9] P. Kaushik, V. Kukreja, E. Jain, M. Ati, and S. Hariharan, "Oral tumor detection and localization using detection transformer (DETR) with multi-scale feature extraction and self-supervised learning," *Array*, vol. 30, p. 100825, 2026, doi: 10.1016/j.array.2026.100825.

[10] C. Karmakar *et al.*, "Hierarchical lesion-aware transformer for oral cancer image classification," *Discov. Artif. Intell.*, 2026, doi: 10.1007/s44163-026-01776-1.

[11] T. R. Sikder *et al.*, "XACT-TB an explainable hybrid CNN–swin transformer framework for tuberculosis screening from chest X-ray images," *Discov. Artif. Intell.*, vol. 6, p. 1030, 2026, doi: 10.1007/s44163-026-01509-4.

[12] M. I. Faruk *et al.*, "Channel attention recalibrated fusion of CoAtNet and frozen DINOv2 features for binary chest X ray classification," *Discov. Comput.*, vol. 29, p. 551, 2026, doi: 10.1007/s10791-026-10446-w.

[13] P. D. Madan Kumar *et al.*, "A Smartphone-based Comprehensive Dataset of Annotated Oral Cavity Images for Enhanced Oral Disease Diagnosis," *Sci. Data*, vol. 13, p. 676, 2026, doi: 10.1038/s41597-026-06954-5.