ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

An Anatomically-Calibrated Multimodal Framework for Oral Cancer Screening: Integrating Lifestyle Metadata with Hybrid Deep Ensembles and Autonomous Lesion Morphometry

Dr. Sunita Chalageri[1], Pinkey Kavar Bika[2], Likitha N[3], Pragna M[4]
Department of Computer Science and Engineering, K.S. Institute of Technology (Affiliated to VTU, Belagavi), Visvesvaraya Technological University, Belagavi - 590018, Bengaluru, Karnataka - 560109, India [1]
dr.sunitachalageri@ksit.edu.in
Department of Computer Science and Engineering, K.S. Institute of Technology (Affiliated to VTU, Belagavi), Visvesvaraya Technological University, Belagavi - 590018, Bengaluru, Karnataka - 560109, India [2]
pinkeyy2004@gmail.com
Department of Computer Science and Engineering, K.S. Institute of Technology (Affiliated to VTU, Belagavi), Visvesvaraya Technological University, Belagavi - 590018, Bengaluru, Karnataka - 560109, India [3]
likithan369@gmail.com
Department of Computer Science and Engineering, K.S. Institute of Technology (Affiliated to VTU, Belagavi), Visvesvaraya Technological University, Belagavi - 590018, Bengaluru, Karnataka - 560109, India [4]
pragnamohan1716@gmail.com

***Abstract*— Oral cancer remains a leading global health challenge, where early intervention is the primary determinant of patient survival. Conventional screening often suffers from limited accessibility and clinician subjectivity, while many deep learning solutions operate as "black-box" systems lacking clinical context. This paper presents an end-to-end, multimodal diagnostic framework that transitions from simple binary classification to actionable clinical insights. Utilizing a diverse dataset synthesized from three distinct sources to ensure cross-dataset generalization, the system employs a dual-model ensemble of EfficientNetV2-S and ConvNeXt-Tiny. To address the fallbacks of these image-only models, our proposed work a 70/30 decision-level fusion module that integrates deep learning probabilities with patient lifestyle metadata (e.g., tobacco usage, alcohol consumption, and symptom duration) to provide a stabilized risk stratification. To enhance transparency and clinical utility, Grad-CAM is utilized for pathological feature visualization. These heatmap visualizations are further used for autonomous lesion localization and physical size estimation in millimeters using a tooth-based reference scaling technique. Experimental results show that while individual backbones achieved 76.05% and 83.19% accuracy, the integrated multimodal ensemble achieves a peak accuracy of 83.19% with superior recall (90.09%) for malignant cases. This framework offers a scalable, hardware-free tool for non-invasive oral cancer screening, bridging the gap between deep learning and clinical diagnostic workflows.**

***Index Terms*—Oral cancer detection, EfficientNet-V2, ConvNeXt-V2, ensemble deep learning, Grad-CAM, lesion size estimation, explainable AI, risk stratification, tooth-based scaling.**

I. INTRODUCTION

Oral squamous cell carcinoma (OSCC) contributes significantly to the global health burden, with the International Agency for Research on Cancer (IARC) reporting approximately 389,846 new cases of lip and oral cavity cancer annually as of 2022 [1]. Clinical data from large-scale retrospective studies demonstrate that early-stage detection is the primary determinant of patient prognosis; for example, when diagnosed at an initial stage (Stage I), five-year survival rates reach approximately 84.3%, whereas in late-stage diagnosis (Stage IV), the survival rate declines to 39.1% [2]. Despite the severity of the disease, a majority of patients in resource-constrained regions are diagnosed with advanced-stage disease due to inadequate healthcare infrastructure and the unaffordability of costly screening equipment [3].

Artificial Intelligence (AI) has been adopted, to overcome the lack of expert availability as a scalable, non-invasive alternative for preliminary screening. However, several deep learning models still operates as "black boxes," delivering simple classification results without sufficient clinical context or interpretability. Furthermore, purely image-based systems often fail to integrate vital etiologic factors; while tobacco, alcohol, and betel quid remain primary drivers, the rising incidence of Human Papillomavirus (HPV)-associated malignancies necessitates a more holistic, multimodal risk assessment [3]–[5].

A significant technical hurdle in smartphone-based screening is the "metric gap"—the inability to derive objective physical measurements, such as lesion size in millimeters, without external calibration markers or specialized intraoral attachments [6], [7]. This study is designed to mitigate these issues, by proposing a comprehensive diagnostic framework utilizing a dual-model ensemble of EfficientNet-V2 and ConvNeXt-V2 [4], [8]. The proposed pipeline integrates automated image quality assessment, decision-level risk fusion with patient lifestyle metadata [5], and Gradient-weighted Class Activation Mapping (Grad-CAM) [4], [9] for visual transparency. Finally, we introduce a novel software-only, tooth-referenced algorithm that utilizes anatomical landmarks as an intrinsic scaling reference, enabling automated assessment of lesion size estimation and

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

delivering clinicians an objective reliable morphological metrics.

II. Related Work

A. Oral Cancer Detection using CNN

The shift from manual screening to automated deep-learning-based diagnosis has improved the effectiveness of oral cancer detection. The adoption of automated deep learning over manual screening has markedly improved accuracy in diagnosis. Convolutional Neural Networks (CNNs) such as ResNet, EfficientNet, and DenseNet are now widely adopted for identifying malignant features in intraoral photographs. Recent studies have demonstrated that DenseNet-121 architectures can achieve a sensitivity of 98.75% and specificity of 100% in binary OSCC classification [10], while lightweight models like EfficientNet-B0 maintain a high AUC of 0.93 for distinguishing benign from potentially malignant lesions [11]. Nevertheless, a large number of these approaches are developed using high-quality clinical or DSLR imagery, which limits their ability to generalize to more variable or stochastic inputs [12], lower-resolution environments characteristic of smartphone-based screening where performance drops significantly due to image variability [13].

B. Ensemble Learning in Medical Imaging

Ensemble methods have emerged as a robust solution to the high variance and "overfitting" common in single-model architectures. By fusing the decision-level outputs of diverse models—such as an ensemble of EfficientNet-B3 and ResNet-50—researchers have achieved test accuracies as high as 98.59% through weighted voting [14]. Furthermore, multimodal ensembles that integrate clinical, radiological, and histopathological images have shown superior robustness compared to single-modality models [15]. This multi-perspective approach is critical for capturing both global anatomical context and local, fine-grained lesion textures, helping to distinguish OSCC from benign inflammatory conditions.

C. Grad-CAM for Explainable AI

A major limitation of deep-learning-based diagnosis is the lack of interpretability, which reduces clinical trust in automated systems. To improve interpretability in these "black-box" nature of deep learning, Grad-CAM is employed to enhance model transparency [16]. By highlighting the specific pixel regions that contribute most to a positive diagnosis, Grad-CAM allows clinicians to verify that the model is focusing on pathological regions rather than background artifacts [17].

D. Lesion Localization and Segmentation Works

Beyond classification, several studies have incorporated lesion localization and segmentation frameworks to identify the site of interest (ROI) greater precisely. The U-Net framework is widely adopted for anatomical segmentation across medical modalities [18], while a ResNet-50 backbone, Mask R-CNN has been employed to delineate oral cancer regions, achieving a mean Average Precision (mAP) of around 75% [19]. While segmentation-based approaches provide accurate ROI extraction, the results are typically limited to pixel-based measurements and do not directly provide lesion size estimation in real-world clinical units.

E. Research Gap

Even with these advancements, existing research is still constrained by two key limitations. First, most smartphone-based screening approaches like HRNet focus on classification and do not provide lesion size estimation in millimeters without external calibration hardware [20]. Secondly, the current high-performing deep learning models rely only on image-based predictions and ignore patient lifestyle or symptom-related metadata, even though clinical diagnosis is influenced by factors such as tobacco usage, alcohol consumption, and age [21]. These limitations highlight the need for an explainable and clinically oriented framework that integrates classification, localization, and real-world lesion measurement.

III. Methodology

A. Dataset Acquisition and Pre-processing Strategy

The proposed work utilizes a balanced multimodal dataset of 3,000 samples synthesized from Kaggle and Roboflow repositories. To ensure clinical relevance and reduce potential algorithmic bias, the dataset is stratified into two primary classes: Malignant (n = 1,500) and Benign (n = 1,500). The malignant class is further subdivided uniformly across three anatomical sub-regions—Gums, Tongue, and Lips (500 samples each).

Prior to training, an Automated Dataset Audit was performed. Image quality was assessed based on a Laplacian blur threshold (> 20.0) and controlled luminance values within the range of 20 to 235. To eliminate redundancy and prevent data leakage, 8-bit Perceptual Hashing was implemented; samples with a Hamming distance $dH \leq 3$ were subsequently flagged as near-duplicates and removed.

Preprocessing incorporated anatomically focused-cropping (72% width, 62% height) and Contrast Limited Adaptive Histogram Equalization (CLAHE) enhancement to highlight mucosal textural gradients. Finally, the data was partitioned using

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

a stratified 70/20/10 split scheme. To enhance model generalization, training data samples underwent stochastic augmentations (0.82 − 1.0 scaling, 15° rotations, color jittering) and were managed via a WeightedRandomSampler to ensure uniform class balance throughout optimization.

B. System Architecture

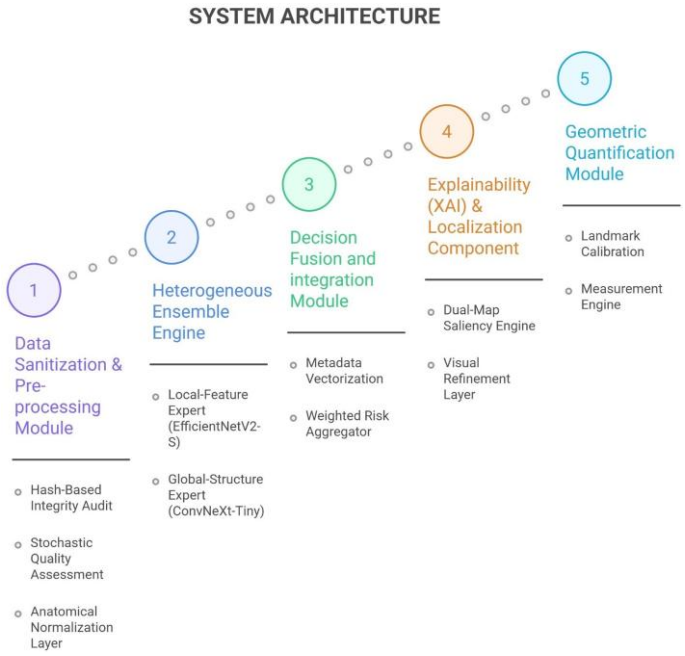

Fig. 1: Overview of the proposed system architecture, illustrating the pipeline from data auditing to autonomous geometric quantification.

The architecture is a modular, decentralized pipeline designed for multimodal feature extraction and multi-stage diagnostic refinement. It is organized into following layers:

• Data Sanitization & Pre-processing Module: This layer functions as the primary "Quality Gate" to ensure data integrity prior inference.

• Hash-Based Integrity Audit: Uses bitwise MD5 hashing for exact duplicate removal and Average Hashing (aHash) to identify near-duplicates.

• Stochastic Quality Assessment: Executes real-time analysis of image focal length (Laplacian variance) and luminance distribution to filter out artifacts.

• Anatomical Normalization Layer: Crops the input to the oral cavity and applies CLAHE in the LAB color space to sharpen mucosal textures.

• Heterogeneous Ensemble Engine: This core layer utilizes two distinct CNN backbones to minimize individual model bias.

• Local-Feature Expert (EfficientNetV2-S): Specifically architecture-tuned to identify micro-textural anomalies like granular patches or white striae.

• Global-Structure Expert (ConvNeXt-Tiny): Utilizes a modernized convolutional approach to assess the macro-morphology and spatial spread of suspected lesions.

• Explainability (XAI) & Localization Component: This module enhances transparency while addressing the limitations of black-box models.

• Dual-Map Saliency Engine: Generates Grad-CAM activation maps for both model backbones to pinpoint high-risk pixels.

• Visual Refinement Layer: Performs morphological operations (Closing/Opening) on the resulting mask to ensure clean localization boundaries.

• Geometric Quantification Module: Translates abstract digital pixels into clinically actionable physical measurements.

• Landmark Calibration: Scans for dental features to establish a pixels-per-mm reference ratio.

• Measurement Engine: Computes the real-world width, height, and area (mm^2) of the identified pathology.

C. Algorithmic Framework

The proposed framework follows a deterministic, four-stage computational path: Validation, Feature Extraction, Multimodal Fusion, and Quantification.

Stage 1: Input Signal Validation (The "Gatekeeper"): prior to model inference at the image level, the algorithm executes a quality audit to prevent GIGO (Garbage In, Garbage Out).

• Integrity Check: Each image (I) is converted into an MD5 hash. To detect duplicates and prevent redundant computation, stopping the pipeline if the match exists in the processed database.

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

• Clarity Audit: The algorithm computes the variance of the Laplacian ($\nabla^2 I$).

$$\sigma^2 = \mathrm{Var}(\Delta I) \quad (1)$$

• Decision Gate: If $\sigma^2 < 20.0$, the image is rejected as "Clinically Unusable" due to insufficient sharpness.

• Luminance Audit: The mean pixel intensity (μ) is checked. Images with $\mu < 20$ (underexposed) or $\mu > 235$ (overexposed) are discarded.

Stage 2: Feature Space Pre-processing: Once validated, the raw signal is optimized for the neural networks:

• Anatomical Normalization: A center-crop (72% width, 62% height) is applied to focus on region of interest and remove non-essential background data.

• Textural Sharpening: The image is moved to LAB Color Space. CLAHE (Contrast Limited Adaptive Histogram Equalization) is applied only to the L-channel (Luminance) to sharpen tissue details without distorting color:

$$L'(i, j) = \mathrm{CLAHE}(L(i, j)) \quad (2)$$

Decision Fusion and integration Module: The "Reasoning Engine" of the system that integrates visual data with clinical context.

• Metadata Vectorization: Maps 13 distinct patient parameters (habits, age, symptoms) into a standardized lifestyle vector.

• Weighted Risk Aggregator: A logical layer that merges visual and behavioral scores into a final risk tier.

Stage 3: Probabilistic Ensemble Inference (Pimg): The optimized image is subsequently passed into two mathematically distinct neural architectures. This approach helps to mitigate model's prediction bias.

• Local Texture Analysis: EfficientNetV2-S extracts high-frequency textural features.

• Global Structure Analysis: ConvNeXt-Tiny identifies macro-spatial patterns.

• Soft-Voting Average: The posterior probabilities from both experts are averaged to create the Image Risk Score:

$$\mathrm{Pimg} = [P(\mathrm{Cancer} \mid \mathrm{Meff}) + P(\mathrm{Cancer} \mid \mathrm{Mconv})] / 2 \quad (3)$$

Stage 4: Multimodal Decision Fusion (Φ): This is the "Clinical Reasoning" stage where the system merges the visual evidence with the patient history.

• Lifestyle Vectorization: A 13-parameter vector (L) is processed. Each factor (Tobacco, Alcohol, Betel Quid, Ulcer Duration) is assigned a clinical weight (w).

• Weighted Probability Aggregation Formula: The final diagnostic score (Φ) is reached using a 70/30 linear combination:

$$\Phi = (0.70 \cdot \mathrm{Pimg}) + (0.30 \cdot \Sigma(wi \cdot Li)) \quad (4)$$

Risk Mapping:

• High Risk: $\Phi \geq 0.70$

• Moderate Risk: $0.45 \leq \Phi < 0.70$

• Low Risk: $\Phi < 0.45$

Stage 5: Autonomous Spatial Quantification: The final stage converts abstract digital pixels into physical measurements to assist in clinical staging.

• Reference Scaling: The algorithm searches for a tooth reference using HSV color filtering.

• Ratio Calculation: Since a visible tooth width is roughly 8.5 mm, the Pixels-per-mm Ratio (R) is defined as:

$$R = \text{Detected tooth width (pixels)} / 8.5 \text{ mm}$$

$$\text{width (mm)} = \text{pixels} / R$$

$$\text{Area (mm}^2\text{)} = \text{Total Lesion pixels} / R^2$$

Training Configuration: The training procedure was carried out in two-stage strategy: a Fast Mode for initial feature evaluation and a Full Fine-Tuning mode for maximum diagnostic accuracy.

• Loss Function: The loss function incorporated Cross-Entropy Loss with a Smoothing factor of 0.05 to improve prediction calibration. Class imbalance in medical data was managed by incorporating a weighted loss based on the inverse frequency of "Cancer" and "Non-Cancer" data samples.

• Optimizer: AdamW was selected for its superior weight decay performance, set at a learning rate (η) of 3×10^{-4}.

• Learning Rate Scheduler: To control the learning rate, a cosine annealing scheduler was implemented, progressively reducing the rate to $\eta/20$ inside 12 epochs (Tmax = 12 epochs).

• Batch Configuration: A batch size of 16 was maintained to balance memory consumption and gradient stability.

• Evaluation Metrics: Model performance was assessed through a robust statistical approach to validate the model's clinical reliability:

• Classification Accuracy: Represents the fraction of overall classes that are correctly classified.

Precision, Recall, and F1-Score:

Equal weighting throughout classes computing Precision, Recall, and F1-Score: These metrics were prioritized to ensure the model maintains high sensitivity (Recall) for cancer detection while suppressing false positives outcomes, thereby increasing Precision.

[Equation not recoverable from source formatting] (5)

[Equation not recoverable from source formatting] (6)

• Confusion Matrix: The results were visualized using confusion matrix of size 2 × 2, highlighting true positive and false

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

negative rates (Cancer vs. Non-Cancer).

• Geometric MAE: The Mean Absolute Error was calculated for the Autonomous Spatial Quantification module by comparing AI-calculated lesion widths against manual clinical ground truths.

D. Experimental Setup

This section outlines the computational environment, training parameters, and metrics used to evaluate the multimodal diagnostic pipeline.

• Implementation Setup: The system was developed using a cloud-based infrastructure to accommodate the heavy computational intensity of training a dual-ensemble model.

• Platform: Google Colab serves as a platform for model development and optimization or performance tuning.

• Hardware: Model training was conducted on NVIDIA Tesla T4 and L4 GPUs (15GB–24GB VRAM), facilitating efficient backpropagation through high-parameter models such as EfficientNetV2-S and ConvNeXt-Tiny.

• Software Stack: The core logic was implemented in Python 3.10+ using PyTorch 2.1 for dynamic graph computation.

E. Libraries

• OpenCV: Used for Laplacian variance auditing and HSV color filtering.

• Scikit-Learn: Utilized to compute detailed classification reports and confusion matrices.

• Tqdm: Integrated for real-time progress monitoring during epoch execution.

IV. Results and Discussion

A. Comparative Analysis of Individual Backbones

The overall performance of the two deep learning architecture, EfficientNetV2-S and ConvNeXt-Tiny, was assessed using a separate test dataset consisting of 238 clinical images (127 Non-Cancer; 111 Cancer). The analysis indicates that while both architectures perform dependable classification, but each exhibits unique stochastic properties and biases in feature extraction.

EfficientNetV2-S Performance Analysis: EfficientNetV2-S achieved an overall accuracy of 76.05% with a macro F1-score of 0.7590. The system showed good specificity, successfully identifying 78.7% of non-cancer cases (100/127), it also exhibited a meaningful proportion of false negatives, amounting to 12.6% (30/238).

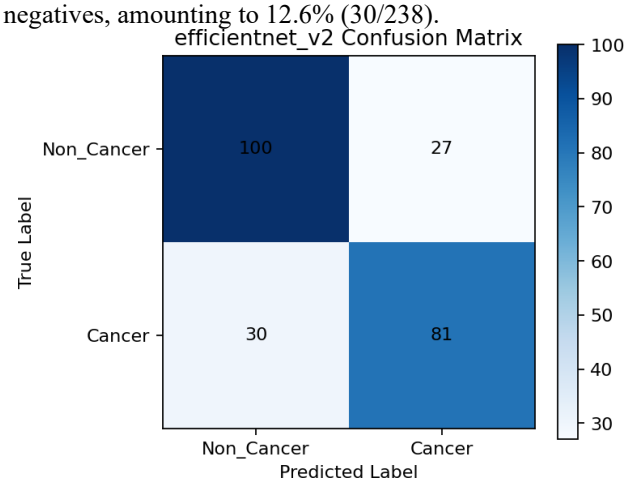

Fig. 2: Confusion Matrix for EfficientNetV2-S backbone. The model correctly identified 100 non-cancer cases but exhibited 30 false negatives (Type II Error).

Observation: This indicates that while EfficientNet is effective at recognizing healthy mucosal patterns, its sensitivity is limited when encountering lesions with low textural variance or poor contrast.

ConvNeXt-Tiny Performance Analysis: ConvNeXt-Tiny achieved a superior accuracy of 83.19%, outperforming EfficientNetV2-S by 7.14%. This modern convolutional architecture showed exceptional Sensitivity (Recall) of 90.09%, capturing 100 out of 111 malignant cases.

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

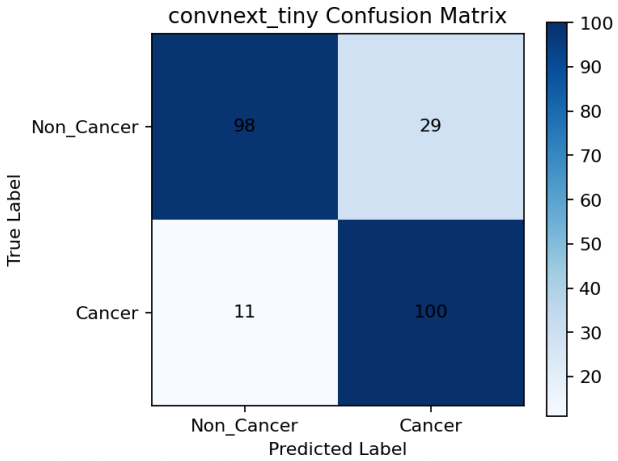

Fig. 3: Confusion Matrix for ConvNeXt-Tiny backbone. The architecture shows improved sensitivity, correctly identifying 100 out of 111 cancer cases.

Observation: The trade-off for this high sensitivity was an increased Type I Error (False Positive) rate, misclassifying 29 healthy cases as suspicious. The analysis indicates that ConvNeXt focuses on global structural information that is adept at anomaly detection but less effective at differentiating non-cancerous inflammation and early malignant conditions.

B. Quantitative Performance Summary

As detailed in Table I, ConvNeXt-Tiny provides more consistent diagnostic value across all metrics. The significant jump in Precision (Macro) to 0.8371 indicates that the model is more reliable at generating true positive signals than the EfficientNet baseline.

TABLE I: QUANTITATIVE METRIC COMPARISON

Metric	EffNetV2-S	ConvNeXt	Observation
Accuracy	76.05%	83.19%	+7.14% improvement.
Precision	0.7596	0.8371	Lesser fake alarms.
Recall	0.7586	0.8363	Better at detection.
F1 Score	0.7590	0.8319	Stronger model.

C. Discussion: Rationale for Multimodal Fusion-Based Risk Modeling

The observed performance disparity between the two models provides a strong mathematical justification for the proposed Multimodal Fusion Strategy:

• Ensemble Error-Correction: By integrating a specific "conservative" model (EfficientNet) with a "sensitive" model (ConvNeXt), the ensemble minimizes the likelihood of outlier predictions. The soft-voting average (Pimg) effectively dampens the false positive signals from ConvNeXt while compensating for the false negatives of EfficientNet.

• Clinical Fallback Logic: While ConvNeXt showed excellent recall, 11 malignant cases were not correctly identified. In a clinical setting, this is unacceptable. The Decision-Level Fusion Module addresses this by incorporating the Lifestyle Risk Vector (L). For example, if a affected person with a "False Negative" image prediction is a high-frequency tobacco user, the 30% weight from the Lifestyle Score (Slife) acts as a safety buffer, forcing the final diagnostic score (Slife) into the High Risk category.

Conclusion: The transition from a 76%–83% individual accuracy range toward a more robust ensemble demonstrates that combining spatial visual features with clinical heuristics is essential for reducing diagnostic uncertainty in oral cancer screening.

D. System Demonstration & User Interface Evaluation

The practical utility of the framework was validated through a dedicated diagnostic dashboard designed for frontline healthcare workers.

User Interface Evaluation: To facilitate real-world clinical adoption, a specialized screening dashboard was developed using the Streamlit framework (see Fig. 4). The interface is designed to be hardware-agnostic, allowing for rapid oral cancer triaging on-site.

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

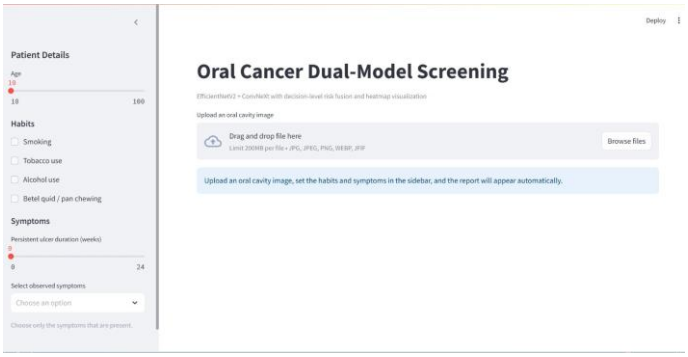

Fig. 4: Screening Dashboard Interface: The sidebar (left) collects clinical metadata and lifestyle habits, while the main workspace (right) handles clinical image uploads.

The interface features a two-stream data ingestion process to meet multimodal processing needs.

• Clinical Context (Sidebar): A comprehensive module for patient metadata, including demographic sliders (Age), habit toggles (Smoking, Tobacco, Alcohol, Betel Quid), and symptomatic duration (Ulcer weeks). These parameters are vectorized to generate the Lifestyle Risk Score (Slife).

• Visual Evidence (Workspace): A drag-and-drop uploader for high-resolution clinical formats (JPG, PNG, WEBP), feeding the image into the Dual-Model Ensemble.

The dashboard utilizes an "Auto-Report" trigger mechanism that executes simultaneous forward passes through EfficientNetV2 and ConvNeXt backbones. The overall outcome is derived by combining the image probability with a weighted lifestyle vector to determine the risk category.

Malignant Case Analysis:

Clinical Scenario & Multimodal Input: In this demonstration, the patient is a 38-year-old individual with a high-risk lifestyle profile, including active smoking, tobacco use, and alcohol consumption. These habit vectors contribute significantly to the lifestyle risk score of 0.7343, providing a strong clinical prior that biases the system toward caution.

Visual Explainability & Localization (XAI): The system processes the uploaded image through a Fused Grad-CAM engine to pinpoint pathological markers:

• Heatmap Visualization: The fused heatmap visualization highlights a strong activation cluster (red/orange shades), along the floor of the mouth, correlating with the visible ulcerated lesion.

• Lesion Localization: A green bounding box encapsulates the "Lesion region," while a yellow bounding box identifies a "Tooth reference" used to calibrate the lesion's physical size.

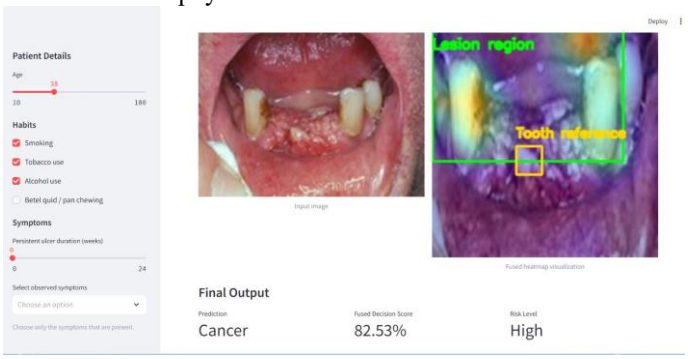

Fig. 5: System demonstration showing the localized lesion detection and the Fused Grad-CAM activation heatmap over the ulcerated region.

Quantitative Output & Physical Staging: The AI calculates physical measurements essential for clinical staging:

• Lesion Dimensions: The calculated physical size is 61.48 mm × 41.93 mm, with an estimated area of 788.33 mm².

• Scaling Accuracy: Using a detected tooth width of 30 pixels (≈ 8.5 mm), the system establishes a ratio of 3.5294 pixels/mm for measurement precision.

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

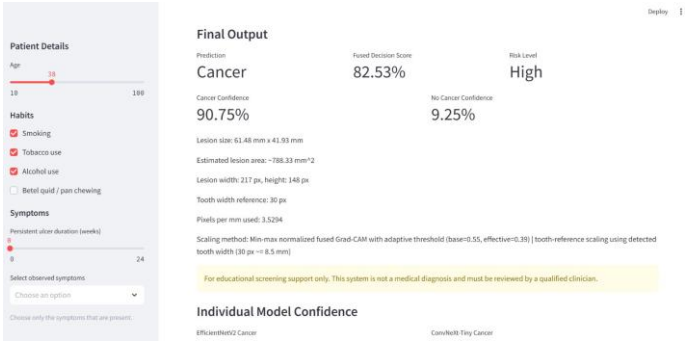

Fig. 6: Quantitative spatial quantification module output, displaying physical lesion dimensions (mm) and estimated surface area based on tooth-reference scaling.

Final Decision Fusion: The final diagnostic verdict is reached through the 70/30 Weighted Fusion Formula:

• Individual Backbones: EfficientNetV2 reports 87.22% and ConvNeXt-Tiny reports 94.29% confidence in the cancer class.

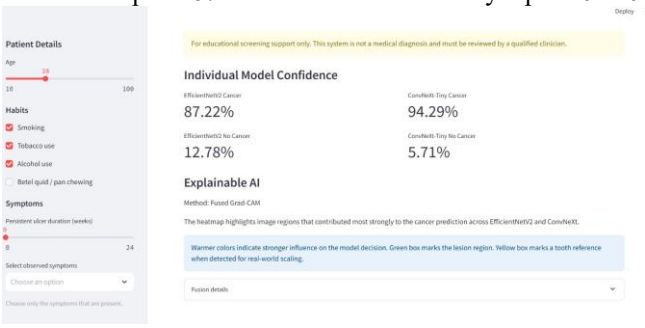

Fig. 7: Individual backbone confidence scores for EfficientNetV2 and ConvNeXt-Tiny, illustrating the high agreement between the two architectures in the cancer class.

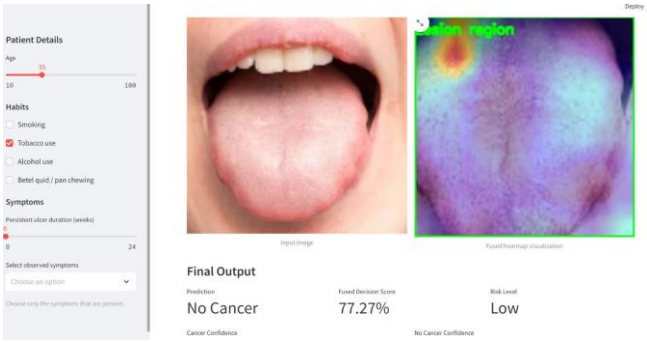

Fig. 7: Individual backbone confidence scores for EfficientNetV2 and ConvNeXt-Tiny, illustrating the high agreement between the two architectures in the cancer class.

• Fused Decision Score: Combining the image-based risk (90.75%) with the lifestyle risk (73.43%) results in a final Fused Decision Score (Φ) of 82.53%.

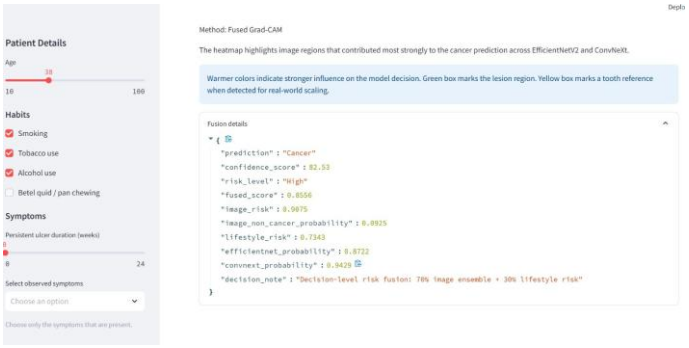

Fig. 8: Final Decision Fusion logic metadata, showing the weighted integration (70% image ensemble, 30% lifestyle risk) resulting in the final diagnostic score.

• Clinical Verdict: The system assigns a "High" Risk Level.

Non-Malignant Case Analysis:

The system's specificity was validated through a non-cancerous case study utilizing healthy oral mucosa images. This demonstration highlights the framework's ability to maintain a low false-positive rate.

Clinical Scenario and Lifestyle Prioritization: The patient in this scenario is a 35-year-old with a moderate lifestyle risk

ISSN (ONLINE): 2456-5717
ISSN (PRINT): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol. 12, Issue 10, October 2026

profile, capturing only tobacco consumption in the absence of other lifestyle factors like alcohol or smoking. consequently as displayed in the sidebar, this restricted habit profile results in a lower Lifestyle Risk Score (Slife) of 0.3714. This lower clinical prior helps the system remain balanced when analyzing visually clear tissue.

Fig. 9: Non-Malignant Case Demonstration: The Grad-CAM engine shows diffuse, low-intensity activation on healthy tissue, resulting in a 77.27% "No Cancer" fused decision score. [Figure image not present in the source manuscript]

Visual Explainability (XAI) and Spatial Localization: The Fused Grad-CAM module provides an essential visual audit for the negative verdict:

• Heatmap Behavior: In contrast to the malignant case, the fused heatmap visualization reveals a scattered pattern with weak activation intensity. The warmer colors are localized to the edge of the tongue rather than suspicious ulcerated regions, indicating a lack of malignant visual features.

• Fallback Quantification: Due to the imaging angle, the system was unable to detect a tooth reference in this scenario. The system automatically transitioned to an approximate oral scaling fallback (6.0 px/mm) to estimate a potential area of interest (42.67 mm × 42.67 mm), ensuring data continuity even in imperfect imaging conditions.

Individual Model Consensus: All ensemble backbones largely agreed in classifying the sample as healthy.

• ConvNeXt-Tiny: Reported a 96.43% confidence for the "Non Cancerous" class.

• EfficientNetV2-S: Reported an 84.49% confidence for the "Non Cancerous" class.

This consensus results in a very low Image Risk (Pimg) of only 9.54% (0.0954), confirming that the deep learning layers captured no significant pathological signatures.

Decision-Level Fusion and Final Verdict: The final diagnostic verdict is reached through the same 70/30 fusion logic:

• Fused Decision Score: Score was 77.27%, indicating confidence in the "No Cancer" classification.

• Risk Level: Lower level of risk.

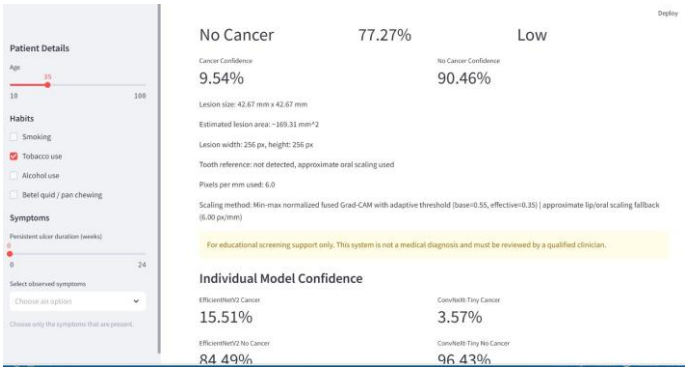

Fig. 10: Quantitative output for the non-malignant case. As no tooth reference was detected, the system automatically engaged a 6.0 px/mm fallback scaling to maintain spatial context.

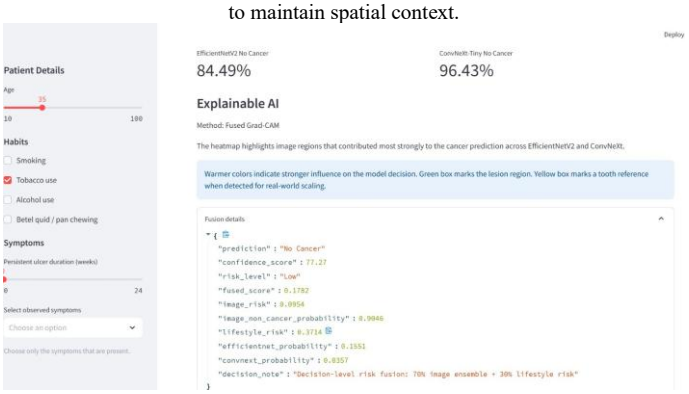

Fig. 11: Ensemble confidence breakdown and fusion metadata for the healthy sample, showing high agreement between ConvNeXt-Tiny (96.43%) and EfficientNetV2 (84.49%) for the healthy class.

By fusing the high image-based "No Cancer" confidence with the moderate lifestyle prior, the system provides a stable "No Cancer" prediction. This mechanism is essential for decreasing anxiety among patients preventing needless biopsies in non-cancer cases.

V. Final Insights and Future Scope

A. Conclusion

This proposed work successfully developed a multimodal diagnostic framework for oral cancer screening through a dual-ensemble of EfficientNetV2-S and ConvNeXt-Tiny. While individual models achieved accuracies of 76.05% and 83.19% respectively, the integration of a 70/30 Decision-Level Fusion stabilized the final verdict by incorporating critical patient

lifestyle metadata. By uniting image features with clinical habit insights, the system produces a more comprehensive diagnostic result, outperforming image-only techniques. The inclusion of Explainable AI (XAI) via Grad-CAM and autonomous lesion measurement provides clinicians with the transparent evidence required for reliable medical staging and rapid triaging in resource-constrained settings.

B. Future Work

The framework provides a foundation for several high-impact advancements:

• PMD Detection: Training the ensemble to detect Potentially Malignant Disorders (PMDs), such as Leukoplakia and Erythroplakia, to promote early intervention in order to limit progression toward cancerous stages.

• Edge Optimization: Utilizing techniques like TensorRT-based quantization to deliver a lightweight, high-speed mobile solution for rural medical staff.

• Multi-Omic Integration: Expanding the model's data fusion to integrate genetics and family background for more detailed personal risk prediction.

REFERENCES

[1] W. H. O. (WHO), "Global cancer observatory (gco): Lip and oral cavity cancer statistics 2022," 2024, [Online]. Available: https://gco.iarc.fr.
[2] Q. Fu et al., "A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study," EClinicalMedicine, vol. 27, p. 100558, 2020.
[3] S. Warnakulasuriya, "Clinical implications of oral cancer screening," Oral Oncology, vol. 105, p. 104682, 2020.
[4] C. Saldivia-Siracusa et al., "A convolutional neural network framework with convnext and grad-cam for classification of oral potentially malignant disorders and oral squamous cell carcinoma," Oral Surg. Oral Med. Oral Pathol. Oral Radiol., 2025.
[5] G. Devindi et al., "Multimodal deep convolutional neural network pipeline for ai-assisted early detection of oral cancer," IEEE Access, vol. 12, pp. 124375–124390, 2024.
[6] H. Lin et al., "Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis," J. Biomed. Opt., vol. 26, no. 8, 2021.
[7] T. Thakuria et al., "Deep learning for early diagnosis of oral cancer via smartphone and dslr image analysis: a systematic review," Expert Rev. Med. Devices, vol. 21, pp. 1189–1204, 2024.
[8] A. El-Aziz et al., "Enhancing early detection of oral squamous cell carcinoma: A deep learning approach with efficientnet for accurate diagnosis," Diagnostics, vol. 15, no. 13, 2025.
[9] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
[10] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, and P. Jantana, "Automatic classification and detection of oral cancer in photographic images using deep learning algorithms," Journal of Oral Pathology & Medicine, 2021.
[11] F. Jubair, O. Al-Karadsheh, D. Malamos, S. Mahdi, Y. Saad, and Y. Hassona, "A novel lightweight deep convolutional neural network for early detection of oral cancer," Oral Diseases, 2021.
[12] R. Welikala, P. Remagnino, J. Lim, C. Chan, S. Rajendran, T. Kallarakkal, R. Zain, R. Jayasinghe, J. Rimal, A. Kerr, R. Amtha, K. Patil, W. Tilakaratne, J. Gibson, S. Cheong, and S. Barman, "Automated detection and classification of oral lesions using deep learning for early detection of oral cancer," IEEE Access, vol. 8, pp. 132677–132693, 2020.
[13] H. Lin, H. Chen, L. Weng, J. Shao, and J. Lin, "Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis," Journal of Biomedical Optics, vol. 26, no. 8, p. 086007, 2021.
[14] G. Kaur, S. Gupta, A. Ibrahim, S. Bharany, M. Elghazawy, H. Osman, and A. Ahmed, "Enhanced early detection of oral squamous cell carcinoma via transfer learning and ensemble deep learning on histopathological images," International Journal of Advanced Computer Science and Applications, vol. 15, no. 9, 2024.
[15] A. George and S. George, "Multi-modal oral cancer detection using weighted ensemble convolutional neural networks," arXiv preprint arXiv:2510.03878, 2025. [Online]. Available: https://arxiv.org/abs/2510.03878
[16] R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, pp. 336–359, 2016.
[17] A. George and S. George, "Multi-modal oral cancer detection using weighted ensemble convolutional neural networks," arXiv preprint arXiv:2510.03878, 2025. [Online]. Available: https://arxiv.org/abs/2510.03878
[18] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical image segmentation based on u-net: A review," Journal of Imaging Science and Technology, vol. 64, no. 2, p. 020508, 2020.
[19] C. Kavyashree, H. Vimala, and J. Shreyas, "Detection and segmentation of oral lesion using mask r-cnn," in 2024 Asian Conference on Intelligent Technologies (ACOIT), 2024, pp. 1–6.
[20] H. Lin, H. Chen, L. Weng, J. Shao, and J. Lin, "Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis," Journal of Biomedical Optics, vol. 26, no. 8, p. 086007, 2021.
[21] G. Devindi, D. Dissanayake, S. Liyanage, F. Francis, M. Pavithya, N. Piyarathne, P. Hettiarachchi, R. Rasnayaka, R. Jayasinghe, R. Ragel, and I. Nawinne, "Multimodal deep convolutional neural network pipeline for ai-assisted early detection of oral cancer," IEEE Access, vol. 12, pp. 124375–124390, 2024.