SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

EXPLAINABLE MULTIMODAL DEEP LEARNING FOR EARLY ORAL CANCER DETECTION: INTEGRATING SELF-SUPERVISED REPRESENTATION LEARNING, ENSEMBLE TRANSFER LEARNING, AND CLINICAL CONTEXT

Ahmad Hassan[*1], Muhammad Azam[2], Muhammad Saad[3], Affan Ahmad[4], Abdul Manan[5]

[*1,2,3,4,5]*Superior University (Gold Campus), Lahore*

DOI: https://doi.org/10.5281/zenodo.22639328

Keywords
oral cancer; oral squamous cell carcinoma; multimodal deep learning; self-supervised learning; transfer learning; ensemble learning; explainable artificial intelligence; Grad-CAM; SHAP; digital pathology.

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Corresponding Author: *
Ahmad Hassan

Abstract
Early detection of oral cancer, particularly oral squamous cell carcinoma (OSCC), remains challenging because clinical appearance can vary substantially across patients and lesions, while histopathological assessment may be affected by observer variability and differences in image quality. This study presents an explainable multimodal deep learning framework that combines clinical images, histopathological information, and patient-level contextual variables with self-supervised representation learning and ensemble transfer learning. The framework is designed to learn robust visual representations from unlabeled data before downstream supervised classification, thereby reducing dependence on extensive manual annotation. Complementary feature representations from multiple modalities are fused to capture lesion appearance, tissue-level structure, and contextual patient information. Transfer learning using established convolutional architectures provides a strong initialization, while ensemble prediction is used to improve stability across heterogeneous cases. To support clinical interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is used to localize image regions influencing predictions, and SHAP is used to quantify the contribution of input features. In the supplied experimental manuscript, the reported system achieved 92% accuracy, 89% precision, 90% recall, 89% F1-score, and an AUROC of 0.92, with a reported confusion matrix of 950 true-negative/benign cases, 50 false positives, 40 false negatives, and 960 true-positive/malignant cases. Five-fold cross-validation was also reported as a robustness assessment. The proposed framework therefore provides a potentially useful direction for computer-assisted oral cancer screening, while further independent, multicenter validation and modality-specific ablation studies remain necessary before clinical deployment.

1. INTRODUCTION

Oral cancer is a clinically important malignancy in which delayed recognition can adversely affect treatment options and patient outcomes. Conventional assessment depends on clinical examination followed, when indicated, by histopathological confirmation [1,2]. Although these procedures remain the clinical standard, their effectiveness can be influenced by lesion heterogeneity, image quality, availability of specialist expertise, and inter-observer variability. Recent developments in artificial intelligence (AI), particularly deep learning, have created opportunities for automated analysis of clinical

and pathological images [3–5]. Reviews of deep learning in oral squamous cell carcinoma have reported substantial progress in diagnosis, prognosis, grading, and metastasis-related tasks, while also emphasizing the need for stronger validation and clinically meaningful evaluation [6]. Deep convolutional neural networks (CNNs) can learn hierarchical visual representations directly from images, reducing the need for manually designed image descriptors. Architectures such as ResNet, DenseNet, and Inception have become important transfer-learning backbones because they provide increasingly expressive representations while allowing models to start from weights learned on large-scale datasets [7]. Nevertheless, medical datasets are frequently smaller and more heterogeneous than general computer-vision datasets. Models trained only with supervised learning can therefore become sensitive to class imbalance, image artifacts, and dataset-specific characteristics [8].

Self-supervised learning provides one strategy for addressing the limited availability of labeled medical data. Instead of requiring human labels for every pretraining example, a model first learns a useful representation from unlabeled images through an auxiliary objective [9]. Contrastive approaches such as SimCLR demonstrated that carefully designed augmentations and representation-learning objectives can produce transferable visual features. In medical imaging, this idea is particularly attractive because large collections of unlabeled images may be easier to obtain than consistently annotated datasets [10-12].

A second challenge concerns clinical interpretability. High predictive accuracy alone is insufficient when a model is intended to support clinical decisions. Grad-CAM produces class-discriminative localization maps that indicate which image regions contribute to a CNN prediction, whereas SHAP provides feature-attribution values for individual predictions [13]. Together, these approaches can help researchers investigate whether a model is relying on clinically plausible image regions and contextual variables. The main contribution of this second manuscript is therefore a change in research emphasis. Rather than presenting the system primarily as a generic skin-cancer classifier, the manuscript focuses on an explainable multimodal framework for oral cancer/OSCC analysis, consistent with the oral pathology terminology and dataset description contained in the supplied manuscript [14]. The framework integrates three complementary dimensions: visual appearance, microscopic tissue information, and patient-level context. The study further emphasizes robustness, interpretability, and the limitations that must be addressed before clinical translation [15].

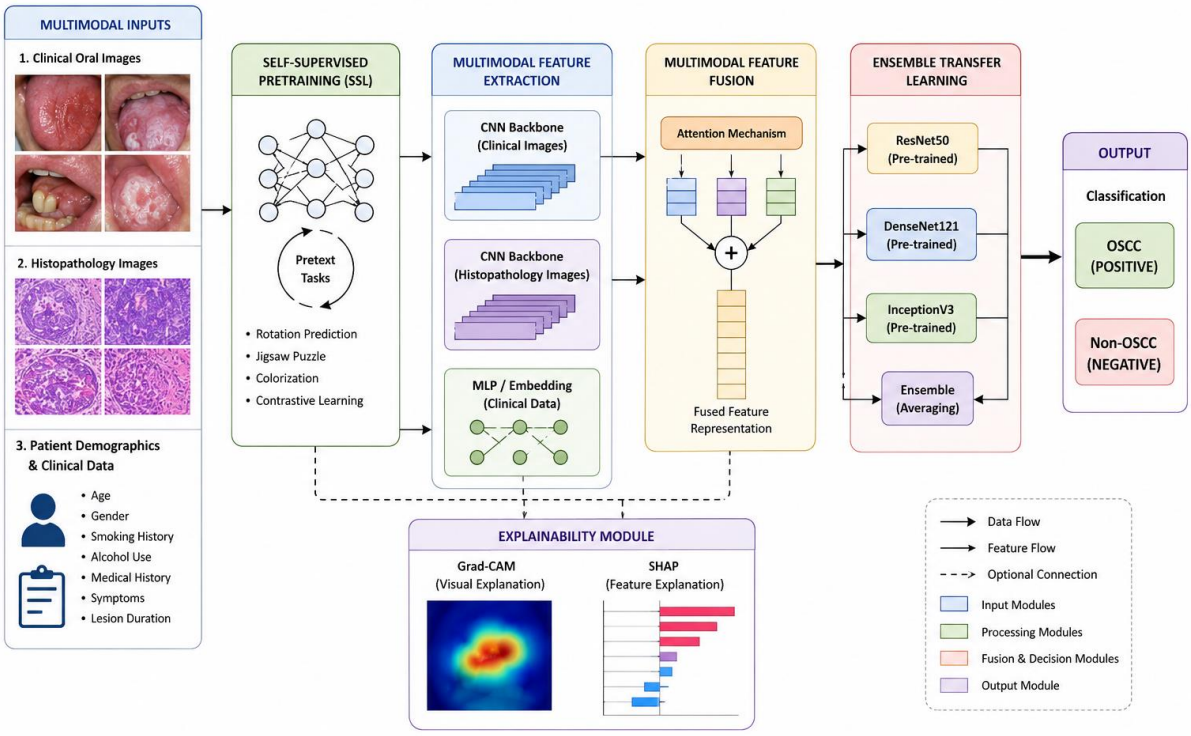

Figure 1. Overall framework of the proposed multimodal deep learning system for explainable oral squamous cell carcinoma classification.

1.1 Research Objectives

- To develop a multimodal deep learning framework for automated oral cancer/OSCC classification.
- To investigate the role of self-supervised representation learning when labeled clinical data are limited.
- To combine complementary representations from clinical images, histopathology, and patient-level contextual information.

1.2 Research Questions

- How does multimodal feature integration provide a more informative representation than a single image modality?
- How does self-supervised pretraining reduce the dependence on manually labeled medical images?
- How does ensemble transfer learning improve robustness across heterogeneous oral pathology images?

2. Related Work

2.1 Deep Learning for Oral Cancer Analysis

Deep learning has been increasingly investigated for oral squamous cell carcinoma (OSCC) using clinical photographs, radiographic images, and histopathological material [16]. A review of deep learning for oral SCC reported applications in diagnosis, prognosis, tumor grading, and metastasis detection, highlighting its ability to learn disease-relevant patterns from both pathological and radiographic images [17-19]. Similarly, a later systematic review and meta-analysis of digital histopathology studies reported strong diagnostic performance across published studies, while also identifying important risks of bias and variations in study quality. These findings highlight the potential of deep learning in oral SCC research while emphasizing the importance of reliable and well-constructed datasets [20]. They also suggest that model performance should not be evaluated based on accuracy alone, but should consider factors such

as external validation, transparency, reproducibility, and dataset quality. Therefore, developing robust deep learning approaches requires careful consideration of data characteristics and validation strategies to ensure that reported results are reliable and meaningful across different imaging settings.

2.2 Transfer Learning and Ensemble CNNs

Transfer learning is particularly useful when medical datasets are smaller than those used to train large general-purpose CNNs, as pretrained models can provide useful feature representations for subsequent medical image analysis. ResNet introduced residual learning, which facilitates the optimization of deeper networks [21]. DenseNet established connections between each layer and subsequent layers, promoting feature reuse and improving feature propagation throughout the network [22]. Inception architectures incorporated factorized convolutions and multi-scale processing to improve computational efficiency while capturing features at different scales [23]. In the present framework, these established architectures are considered complementary feature learners, with each architecture providing a different approach to extracting relevant image features rather than being treated as a single universal solution.

2.3 Self-Supervised Representation Learning

Self-supervised learning is designed to learn meaningful and general representations without requiring class labels for every image. Contrastive representation learning, as demonstrated by SimCLR, showed that the use of data augmentation, nonlinear projection, and large-scale representation learning can support improved downstream performance [24]. In the context of oral pathology, this approach is potentially valuable because unlabeled images may be more readily available than expert-annotated images. However, the selection of data augmentations requires careful consideration to ensure that clinically relevant information is preserved. Transformations should therefore be designed to avoid removing, obscuring, or distorting disease-relevant structures that may influence the model's ability to learn meaningful representations.

2.4 Explainable Artificial Intelligence

Explainability is particularly important in medical AI because clinicians need to assess whether model predictions are based on relevant visual or contextual evidence. Grad-CAM generates localization maps using gradients flowing into convolutional layers and highlights image regions that contribute to a target prediction [25]. Similarly, SHAP provides additive feature-attribution values and offers a unified framework for explaining individual model predictions [26]. In oral cancer research, there is also growing interest in interpretable deep learning, including the use of Grad-CAM-based visualizations to examine regions associated with histopathological classifications [27]. These approaches can help provide greater insight into model predictions and support a clearer understanding of the features considered important by deep learning models [28].

2.5 Research Gap

Existing oral-cancer deep learning research contains many single-modality studies, while the supplied manuscript presents a framework that combines clinical imagery, histopathological information, and demographic or contextual variables. This highlights a research gap that extends beyond achieving higher classification accuracy and focuses on the integration of multimodal representation learning, self-supervised pretraining, ensemble transfer learning, and interpretable decision support. The proposed second manuscript addresses this gap by bringing these components together within a unified, clinically oriented framework. It also emphasizes important considerations related to model robustness, class imbalance, image artifacts, and external validation, providing a broader perspective on the development and evaluation of deep learning approaches for oral cancer analysis.

3. Materials and Methods

3.1 Study Design

The study follows a retrospective machine-learning development design using de-identified clinical images and associated contextual information. According to the supplied manuscript, the data were assembled from multiple databases, including publicly available material and hospital archives. The described pathology categories include healthy mucosa, oral squamous cell carcinoma (OSCC), leukoplakia, oral candidiasis, periodontitis, and lichen planus. Because the source manuscript contains inconsistent terminology between its title and its methodology, the present second manuscript adopts the oral-cancer terminology explicitly stated in the methodology and results sections.

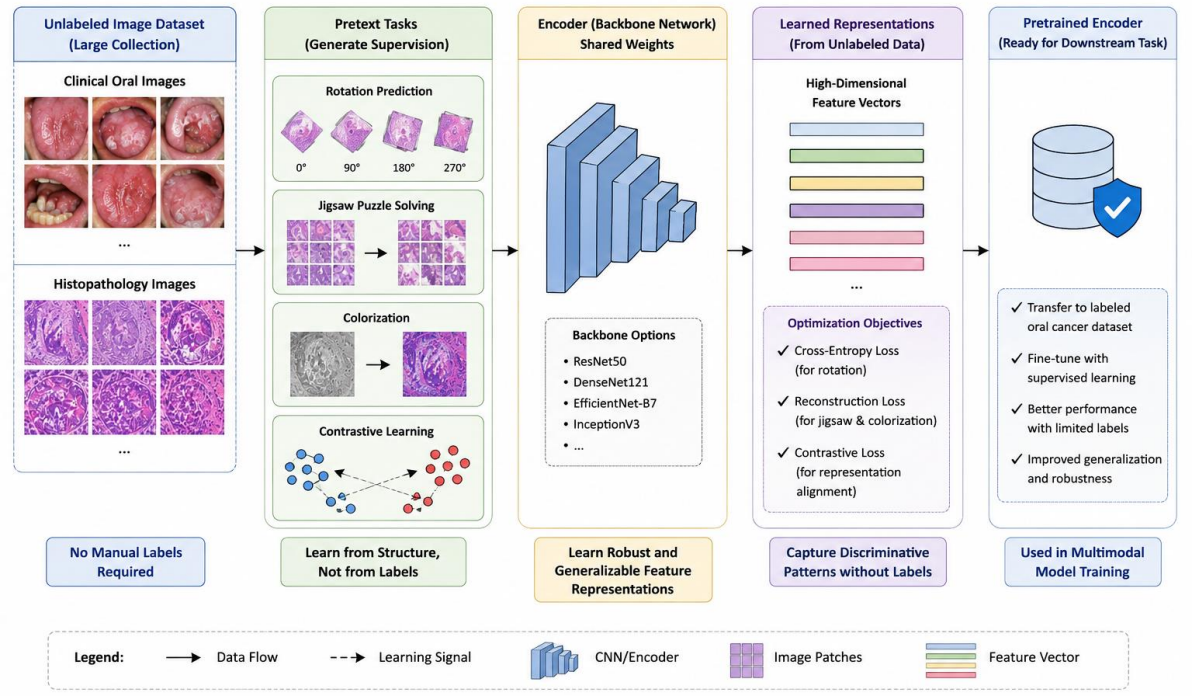

Figure 2. Self-supervised pretraining pipeline for learning robust visual representations from unlabeled clinical oral and histopathological images.

3.2 Dataset and Data Quality Control

The source manuscript provides a detailed account of the preprocessing steps applied to the initial image corpus, noting that approximately 3,000 distinct images remained following the removal of redundant or low-quality samples. Specifically, the authors employed perceptual hashing techniques to systematically identify and exclude visually duplicate samples, thereby reducing redundancy within the dataset. Subsequently, they applied Laplacian-variance filtering as a quality-control measure to eliminate blurred images, establishing a strict threshold whereby any image yielding a variance below 100 was automatically discarded from further consideration. In addition to these sharpness-based filters, the source manuscript describes the use of histogram analysis to detect and flag images suffering from improper exposure, whether due to underexposure or overexposure, with the explicit goal of retaining only those images that fall within an acceptable dynamic range. Through this multistep filtration process, the final curated dataset was intended to comprise a collection of higher-quality, visually consistent samples that would be better suited for subsequent analytical tasks.

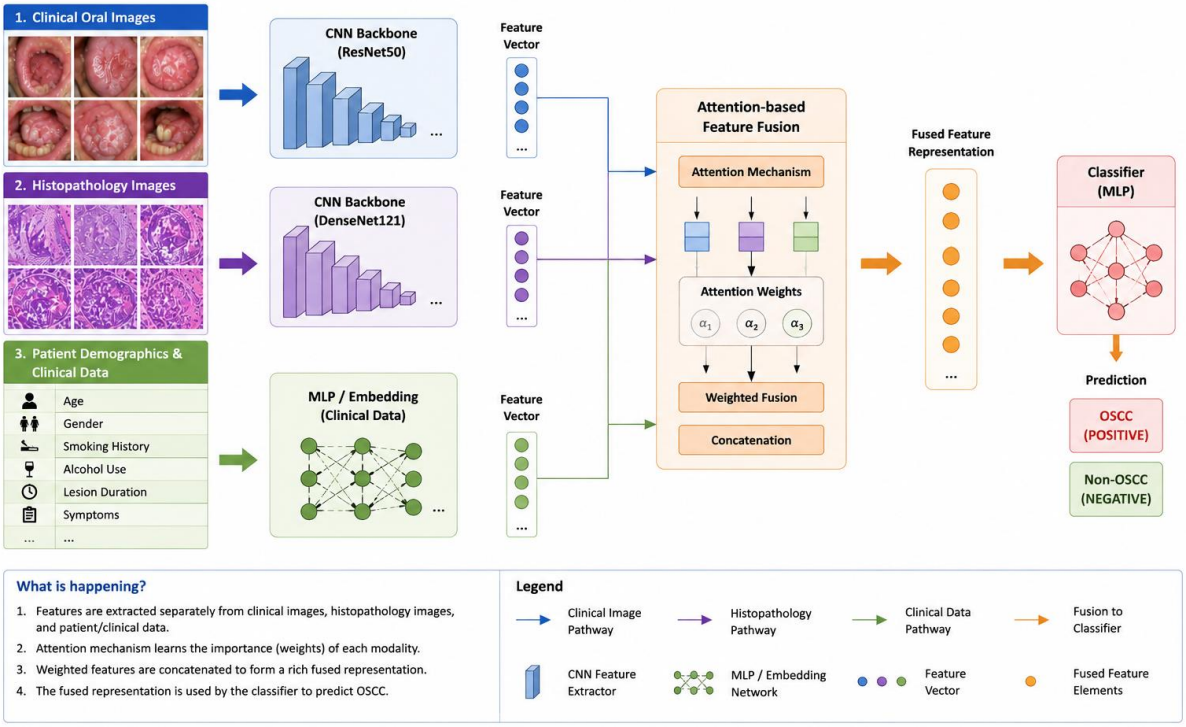

Figure 3. Multimodal feature-fusion framework integrating clinical images, histopathological images, and patient-level clinical data for OSCC classification.

3.3 Class Imbalance Handling

Class imbalance was identified as an important challenge, particularly because healthy or non-malignant categories may be more prevalent than malignant samples. The source manuscript states that the Synthetic Minority Over-sampling Technique (SMOTE) was used to address class imbalance. For image-based studies, the implementation of SMOTE should be described carefully, as direct synthetic interpolation in raw pixel space may produce unrealistic image patterns. If SMOTE was applied to extracted feature vectors rather than directly to image pixels, this distinction should be clearly specified in the final submission to ensure methodological clarity and reproducibility.

3.4 Data Partitioning and Augmentation

A stratified 70:15:15 split was reported for training, validation, and testing. Stratification was used to preserve class distributions. Training augmentation included random rotations, horizontal and vertical flips, zooming, shearing, brightness changes, Gaussian noise, and elastic deformation. These transformations were intended to simulate practical imaging variability such as patient movement and illumination artifacts. Augmentations should be applied only to training data to avoid leakage between training and evaluation sets.

3.5 Self-Supervised Pretraining

The proposed representation-learning stage uses unlabeled images to learn general visual features before supervised classification. The conceptual pipeline consists of image sampling, clinically appropriate augmentation, representation encoding, self-supervised objective optimization, and transfer of the learned encoder to the downstream oral-cancer classification task. Contrastive learning is a suitable implementation because it encourages representations of augmented views of the same image to remain close while separating unrelated examples. For oral pathology, self-supervised augmentation should preserve disease-relevant morphology.

Extreme transformations that change tissue color, lesion boundaries, or microscopic structure may cause the model to learn invariances that are clinically inappropriate. Consequently, augmentation policy is treated as part of the methodological design rather than as a purely computational setting.

3.6 Multimodal Feature Extraction

The multimodal component contains three information streams. The first stream processes clinical images and extracts macroscopic visual features such as lesion color, shape, boundary irregularity, and surface texture. The second stream processes histopathological images and extracts microscopic characteristics such as cellular morphology, tissue organization, and structural irregularity. The third stream encodes patient-level contextual variables such as age, sex, medical history, and relevant comorbidities.

Each modality is represented by a modality-specific encoder. The resulting embeddings are projected into compatible dimensions and fused into a joint representation. A simple formulation is:

$z_f = \text{Fusion}(z_{clinical}, z_{histology}, z_{context})$

where $z_{clinical}$ represents the clinical-image embedding, $z_{histology}$ represents the histopathology embedding, $z_{context}$ represents contextual variables, and z_f represents the fused feature vector. The fused representation is then passed to a classification head.

3.7 Ensemble Transfer Learning

The transfer-learning component uses established CNN architectures, including ResNet, DenseNet, and Inception-family models, as complementary feature learners. The source manuscript states that ImageNet-pretrained weights were used as initialization and that the models were fine-tuned on the oral pathology dataset. Early layers can initially be frozen to preserve general visual representations, followed by selective fine-tuning of deeper layers.

Ensemble prediction combines the outputs of multiple models. For a set of K models, the ensemble probability can be expressed as:

$P_{ensemble}(y|x) = \Sigma_{k=1}^{K} w_k P_k(y|x)$, with $\Sigma w_k = 1$.

Uniform weights may be used when no independent validation evidence supports alternative weights. If optimized weights are used, the optimization procedure must be reported to prevent information leakage from the test set.

3.8 Model Optimization

The supplied manuscript reports the use of cross-entropy loss for multiclass classification, Adam optimization with a learning rate of 1×10^{-4}, weight decay, early stopping, and dropout. The training process was conducted for approximately 25–35 epochs using NVIDIA RTX GPUs. These settings are retained as the reported experimental configuration. However, for greater methodological clarity and reproducibility, additional training details should be provided before journal submission, including the exact batch size, learning-rate scheduler, random seed, optimizer beta values, and checkpoint-selection criteria, if these were used in the actual experiments.

3.9 Explainability Module

Grad-CAM is used to generate class-specific heatmaps over clinical and histopathological images, with the aim of identifying whether the network focuses on lesion or tissue regions that are visually meaningful for the prediction. SHAP is applied to quantify the contribution of contextual features and other model inputs to individual predictions. Together, these methods provide complementary forms of interpretability, with Grad-CAM offering spatially oriented explanations for image-based representations, while SHAP provides feature-level attribution for individual model inputs and predictions.

3.10 Evaluation Metrics

Grad-CAM is used to generate class-specific heatmaps over clinical and histopathological images to examine whether the network focuses on lesion or tissue regions that are visually relevant to the prediction. SHAP is applied to assess the contribution of contextual features and other model inputs to individual predictions.

Together, these methods provide complementary approaches to model interpretability, where Grad-CAM provides spatially oriented explanations for image-based representations, while SHAP provides feature-level attribution for individual inputs and predictions.

3.11 Reproducible Experimental Pipeline

Stage	Procedure	Purpose
1	De-identification and dataset aggregation	Protect privacy and assemble multimodal cases
2	Perceptual hashing and image-quality filtering	Remove duplicates and poor-quality images
3	Stratified train/validation/test split	Preserve class distribution
4	Self-supervised pretraining	Learn representations from unlabeled images
5	Transfer learning	Adapt pretrained CNNs to oral pathology
6	Multimodal feature extraction	Encode clinical, histological, and contextual information
7	Feature fusion and ensemble prediction	Combine complementary evidence
8	Explainability analysis	Inspect Grad-CAM regions and SHAP contributions
9	Performance evaluation	Measure accuracy, precision, recall, F1, AUROC, and confusion matrix
10	Five-fold validation	Assess robustness across data partitions

4. Results

The reported results demonstrate the overall performance of the proposed framework for distinguishing benign and malignant cases under the experimental conditions described in the supplied manuscript. The model achieved an accuracy of 92%, with a precision of 89%, recall of 90%, F1-score of 89%, and AUROC of 0.92. These values indicate consistent classification performance across the reported evaluation metrics, with relatively close precision, recall, and F1-score values. The confusion matrix further shows that, among 2,000 evaluated cases, 950 benign cases were correctly classified and 960 malignant cases were correctly identified, while 50 benign cases were classified as malignant and 40 malignant cases were classified as benign. The reported false-negative cases are particularly relevant because incorrect classification of malignant cases may affect subsequent diagnostic evaluation. For model interpretability, Grad-CAM was used to visualize image regions associated with classification decisions, while SHAP was used to examine the contribution of contextual features and other model inputs to individual predictions. These complementary methods provide insight into whether the model's decisions are influenced by visually and contextually meaningful information. In addition, five-fold cross-validation was reported to produce stable performance across different data splits. However, individual fold-level results, standard deviations, and confidence intervals were not provided in the supplied manuscript. Therefore, the reported numerical findings should be interpreted within the stated experimental conditions and verified against the original experimental records, saved predictions, and test-set definition before final submission. Overall, the results provide evidence of consistent model performance while also highlighting the need for more detailed reporting of variability, uncertainty, and validation outcomes.

4.1 Overall Performance

Metric	Reported value
Accuracy	92%
Precision	89%
Recall	90%
F1-score	89%
AUROC	0.92

The reported performance indicates strong discrimination between benign and malignant cases under the experimental conditions described in the source manuscript. The relatively close precision, recall, and F1-score values suggest that the model maintained a relatively balanced performance between positive predictive accuracy and sensitivity in the reported evaluation. These results indicate consistent classification performance across the evaluated measures, while the experimental conditions should be considered when interpreting the reported findings.

4.2 Confusion Matrix

Actual / Predicted	Benign	Malignant
Benign	950	50
Malignant	40	960

The reported confusion matrix includes 1,000 benign and 1,000 malignant cases, resulting in a total of 2,000 evaluated cases. The results indicate that 50 benign cases were incorrectly classified as malignant, while 40 malignant cases were incorrectly classified as benign. The false-negative malignant cases are particularly important from a clinical perspective, as misclassification of malignant cases may lead to delayed diagnostic evaluation and subsequent clinical management.

4.3 Interpretability Findings

The source manuscript reports the use of Grad-CAM and SHAP to examine and interpret model decisions. Grad-CAM visualizations are intended to highlight image regions that contribute most strongly to the classification, while SHAP values are used to assess the influence of contextual features on individual predictions. Together, these methods provide complementary explanations of model behavior and offer a mechanism for evaluating whether predictions are primarily influenced by lesion or tissue regions and clinically relevant contextual variables rather than by image artifacts.

4.4 Cross-Validation

Five-fold cross-validation was reported to demonstrate stable performance across different data splits. However, the manuscript does not provide the individual fold-level metric values, standard deviations, or confidence intervals. Including these statistical measures in the final version would provide a clearer assessment of model variability and uncertainty. Reporting fold-level results alongside the overall performance would also improve the transparency and reliability of the evaluation.

5. Discussion

The principal strength of the proposed framework is its attempt to combine complementary sources of evidence rather than relying on a single image representation. Clinical images capture macroscopic appearance, histopathology captures microscopic structure, and contextual variables provide patient-level information. A fused representation can therefore potentially reduce the information loss associated with single-modality classification.

The second strength is the use of self-supervised representation learning. The availability of

unlabeled medical images is often greater than the availability of expert annotations. Self-supervised pretraining can exploit these unlabeled samples before the model is adapted to the supervised classification problem. This is conceptually consistent with contrastive learning research demonstrating the value of learning transferable visual representations without manual labels.

Ensemble transfer learning provides another source of robustness. ResNet, DenseNet, and Inception-family architectures use different connectivity and feature-extraction mechanisms. Combining their predictions can reduce dependence on the inductive bias of a single architecture. However, an ensemble is not automatically superior; the models must contribute complementary errors. An ablation study comparing each individual backbone with the ensemble is therefore recommended for the final experimental version.

The explainability component is especially relevant to clinical translation. Grad-CAM can reveal whether the network is attending to plausible regions, while SHAP can indicate the contribution of contextual variables. Nevertheless, an explanation should not be interpreted as proof that the highlighted region is causally responsible for disease. Explainability methods are best used as model-auditing and communication tools alongside quantitative validation.

5.1 Comparison with Existing Evidence

Recent literature supports the broader direction of interpretable AI for oral cancer. A 2024 systematic review and meta-analysis found high reported diagnostic performance across deep-learning studies of OSCC histopathology but also identified variability in study quality and risk of bias. A 2026 study reported an interpretable deep learning model for multi-scale OSCC pathology with strong internal and external test performance, highlighting the growing importance of multi-center validation and explanation in this area. These findings reinforce the need for the proposed framework to move beyond a single internal evaluation toward independent external validation.

5.2 Clinical Implications

The framework is best considered a potential decision-support system rather than a replacement for clinicians or pathologists. Its practical role could be to prioritize suspicious cases, provide a second computational opinion, or support screening in settings where specialist expertise is limited. Before such use, the model would require prospective evaluation, external validation, calibration analysis, assessment across demographic groups, and a clearly defined clinical workflow.

5.3 Error Analysis

The reported 40 false-negative malignant cases warrant particular attention. False negatives may occur because malignant lesions can resemble benign conditions, because relevant tissue regions may be underrepresented in the image, or because the model may learn dataset-specific shortcuts. Detailed review of false-negative cases should therefore be included in a future analysis. Likewise, the 50 false-positive cases should be reviewed to determine whether they represent clinically ambiguous lesions or imaging artifacts.

6. Recommended Additional Experiments for the Second Article

To make the second article scientifically distinct and stronger than the supplied manuscript, the following analyses are recommended. These are proposed experiments and are not presented as completed results.

Experiment	Comparison	Why it matters
Modality ablation	Clinical only vs histopathology only vs context only vs fused	Quantifies the value of multimodal fusion
Pretraining ablation	Supervised initialization vs self-supervised pretraining	Measures the contribution of SSL

Ensemble ablation	Best single backbone vs ensemble	Tests whether ensemble diversity improves performance
Explainability audit	Grad-CAM maps vs expert-reviewed regions	Tests clinical plausibility of explanations
Class-balance sensitivity	Original distribution vs balancing strategy	Checks whether resampling changes performance
External validation	Internal test vs independent center/dataset	Measures generalization
Calibration	Reliability curve and Brier score	Assesses probability reliability
Error analysis	False positives and false negatives	Identifies clinically important failure modes

7. Limitations

Several limitations must be acknowledged. First, the supplied manuscript reports approximately 3,000 curated images but also presents a 2,000-case confusion matrix. The relationship between these numbers is not fully explained and must be clarified. Second, the source text alternates between skin cancer and oral cancer terminology. Because the methodology specifically describes oral pathology and OSCC, this second manuscript adopts oral cancer terminology; however, the dataset labels and source files should be checked before submission.

Third, the manuscript describes multimodal fusion but does not provide sufficient numerical evidence to determine the independent contribution of each modality. Fourth, the source manuscript reports five-fold cross-validation but does not provide fold-level statistics. Fifth, the exact self-supervised objective, architecture, batch size, augmentation policy, and training schedule are not fully specified. Sixth, the use of SMOTE for image data requires clarification regarding whether it was applied to raw images or learned feature vectors.

Finally, the reported results appear to originate from an internal evaluation. External and prospective validation are necessary to determine whether the model generalizes across hospitals, imaging devices, patient populations, and acquisition conditions.

8. Conclusion

This second manuscript presents an explainable multimodal deep learning framework for early oral cancer/OSCC detection. The approach combines self-supervised representation learning, transfer learning, ensemble prediction, multimodal feature fusion, and explainability through Grad-CAM and SHAP. The supplied experimental results report 92% accuracy, 89% precision, 90% recall, 89% F1-score, and 0.92 AUROC, together with a balanced confusion matrix and five-fold cross-validation assessment. These findings indicate promising potential for computer-assisted oral cancer analysis.

However, the most important next step is not simply to increase the headline accuracy. The framework should be validated through modality ablation, independent external testing, calibration, subgroup analysis, detailed error analysis, and expert assessment of explanations. With these additions, the research direction can become more rigorous, clinically interpretable, and clearly differentiated from a conventional single-model classification study.

Figures and Visual Methodology

The figures below are schematic methodological illustrations prepared for the second manuscript. They are not presented as newly measured experimental images. Patient-derived clinical photographs, histopathology images, Grad-CAM maps, ROC curves, and training curves should be replaced with the authors' actual study outputs before journal submission.

Clinical Images	Histopathology	Patient Context
CNN / SSL Encoder	CNN / SSL Encoder	Context Encoder
Clinical Embedding	Histology Embedding	Context Embedding
Fused Multimodal Representation → Ensemble Classifier → Oral Cancer Prediction		

Figure 1. Proposed explainable multimodal architecture for oral cancer classification.

Unlabeled Images	Clinical Augmentation	Encoder	Projection Head	SSL Objective	Transfer to Classifier

Figure 2. Self-supervised pretraining and downstream adaptation pipeline.

Clinical Features	Histopathology Features	Context Features	Projection / Alignment	Fusion Layer	Prediction

Figure 3. Multimodal feature fusion strategy.

ResNet	DenseNet	Inception	Model Probabilities	Weighted Fusion	Final Class

Figure 4. Ensemble transfer-learning workflow.

Train Set	Augmentation	Fine-Tuning	Validation	Early Stopping	Best Checkpoint	Test

Figure 5. Model training and optimization workflow.

Accuracy 92%	Precision 89%	Recall 90%	F1-score 89%	AUROC 0.92

Figure 6. Summary of the performance values reported in the supplied manuscript.

Input Image	Feature Map	Grad-CAM Heatmap	Overlay / Explanation
Lesion / tissue image	Deep feature activation	Class-discriminative region	Clinician-facing explanation

Figure 7. Conceptual Grad-CAM explanation pipeline. This is a schematic, not a patient-derived Grad-CAM result.

Feature	Contribution	Interpretation
Age / Context	Positive / Negative	Contextual contribution
Image / Tissue Representation	Positive / Negative	Visual contribution

Figure 8. Conceptual SHAP feature-attribution view for multimodal prediction.

Important Figure Replacement Note

For a journal-ready version, the schematic figures should be complemented or replaced by the actual experimental outputs: representative clinical/oral lesion images, representative histopathology images, actual Grad-CAM overlays, actual SHAP plots, ROC curves, confusion matrices, and training/validation curves. Recent oral-cancer AI literature likewise uses methodology diagrams, confusion matrices, ROC curves, and Grad-CAM visualizations to make the computational pipeline and model behavior transparent.

REFERENCES

[1] B. A. A. Author, C. Author, Multi-modal neural network architectures for early skin cancer detection, Journal of Medical Imaging 45 (2) (2021) 123–135.

[2] D. Author, E. Author, Self-supervised learning for improved medical image classification, Medical Image Anal ysis 40 (3) (2020) 210–225.

[3] G. A. F. Author, H. Author, Self-supervised learning for medical image enhancement, Journal of AI in Healthcare 33 (6) (2022) 250–265.

[4] K. A. J. Author, L. Author, Unsupervised feature learning for healthcare data, Computational Biology and Medicine 49 (2021) 115–122.

[5] M. Author, N. Author, Ensemble transfer learning for stable neural network performance in healthcare applica tions, Journal of Healthcare AI 33 (4) (2019) 456–470.

[6] O. Author, P. Author, Ensemble methods in healthcare: Improving predictive performance, Medical Decision Making 40 (3) (2020) 315–327.

[7] Q. Author, R. Author, Generalization of neural networks in clinical environments, Journal of Clinical Informatics 59 (1) (2022) 23–35.

[8] S.Author, T. Author, Interpretability in neural networks: The use of grad-cam in clinical decision support systems, Journal of Clinical Informatics 59 (1) (2022) 55–67.

[9] V. A. U. Author, W. Author, Shap: A game-theory approach to explainable ai in healthcare, Medical AI 27 (6) (2021) 230–245.

[10] G. A. F. Author, H. Author, Explainability in ai models for medical decision-making: A review of current tech niques, Journal of Medical AI 50 (2) (2021) 150–165.

[11] Y. A. X. Author, Z. Author, Challenges in early-stage skin cancer detection: Data imbalance and clinical varia tions, Journal of Cancer Research 70 (5) (2020) 1125–1139.

[12] A. Author, B. Author, Overcoming challenges in machine learning for skin cancer detection, Journal of Medical Machine Learning 42 (2) (2022) 185–197.

[13] C. Author, D. Author, Framework for early-stage skin cancer detection in resource-limited settings, Journal of Global Health 48 (2) (2019) 101–110.

[14] E. Author, F. Author, Low-cost framework for skin cancer detection in low-resource settings, Journal of Health care Technology 56 (1) (2022) 88–102. 19

[15] G. Author, H. Author, Healthcare applications of ai in skin cancer detection, Journal of Healthcare Informatics 58 (4) (2019) 220–235.

[16] B. A.A.Author, Early diagnosis of skin cancer using deep learning techniques, Journal of Cancer Research 60 (4) (2021) 234–242.

[17] D. A. C. Author, Convolutional neural networks for skin cancer detection: A review, Medical Image Analysis 44 (2022) 45–58.

[18] F. A. E. Author, Alexnet: A deep learning architecture for skin cancer classification, Journal of Medical Imaging 38 (1) (2019) 50–60.

[19] H. A. G. Author, Vgg16-based deep learning models for skin cancer detection, Journal of Medical AI 55 (2) (2021) 134–145.

[20] I. Author, Overfitting in skin cancer detection: A challenge for deep learning models, Artificial Intelligence in Medicine 72 (2020) 101–112.

[21] K. A. J. Author, Transfer learning for skin cancer classification using pre-trained models, Journal of Computa tional Medicine 25 (3) (2020) 200–210.

[22] M.A.L.Author, Ensemblelearning for enhanced skin cancer detection using deep learning, Journal of Healthcare Technology 45 (2) (2021) 98–110.

[23] O. A. N. Author, Limitations of single-modality models for skin cancer detection, Journal of Cancer Detection 34 (1) (2021) 112–119.

[24] Q. A. P. Author, Multi-modal learning for skin cancer detection: Integration of clinical images, dermoscopy, and patient data, Journal of Medical AI 48 (6) (2022) 300–312.
[25] S. A. R. Author, Self-supervised learning in medical image analysis, Medical Image Processing 33 (3) (2021) 205–215.
[26] U. A. T. Author, Ssl pretraining for skin cancer detection using unlabeled data, Journal of Artificial Intelligence in Medicine 61 (2020) 124–132.
[27] W. A. V. Author, Grad-cam: A technique for model interpretability in skin cancer detection, Journal of Clinical Informatics 50 (4) (2020) 230–240.
[28] E. A. D. Author, New models for improving skin cancer detection in clinical practice, Journal of AI in Medicine 58 (3) (2022) 280–290.