



RESEARCH ARTICLE

Multimodal Generative Reconstruction of Oral Cancer Lesions From Sparse 2D Data Using Attention-based Deep Learning

Abhilash Shukla, Dhatri Raval, Kalpit Soni, Atul Patel

Abstract

Early and proper diagnosis of oral cancer is essential in improving patient survival and giving an opportunity to make clinical interventions. However, the acquisition of good quality three-dimensional (3D) diagnostic information remains a major issue due to its inaccessibility, exorbitant cost, and the requirement of special imaging equipment. Clinicians are therefore inclined to use limited two-dimensional (2D) information including intraoral images and histopathological slide images that might not capture the intricate structural features of the malignant lesions. To overcome these shortcomings, the current paper suggests a new template of the generative reconstruction of oral lesions associated with cancer based on sparse and multimodal 2D images. The proposed model is based on the strength of convolutional neural networks to find strong features along with an attention-based fusion mechanism to successfully incorporate the heterogeneous data sources. The integration enables the generation of a uniform and enhanced image of the lesions without any loss of critical morphological information which is necessary to come up with an adequate diagnosis. The model is capable of learning complex spatial structures, inter-modal interactions and re-assembling very fine lesion shapes with very low data requirements. The quality of reconstruction utilizing the proposed method has been demonstrated to be much higher and affect the task of downstream cancer classification. The Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) quantitative analysis is also successful in demonstrating that the framework is even superior to the traditional reconstruction techniques. This research broadly provides a high-scaling, low-cost tool to improve the accuracy of the diagnosis in the environment with scarce resources. The suggested framework is a hopeful outlook towards the early identification and proper treatment of oral cancer by removing the necessity to use sophisticated imaging modalities, yet securing clinical credibility.

Keywords: Oral Cancer, Multimodal Medical Imaging, Deep Learning, Attention-Based Fusion, Generative Reconstruction

Introduction

Oral cancer has been one of the health challenges facing the population of the world with high morbidity and

Faculty of Computer Science and Applications (FCA), Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India.

***Corresponding Author:** Dhatri Raval, Faculty of Computer Science and Applications (FCA), Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India., E-Mail: dhatriraval.mca@charusat.ac.in

How to cite this article: Shukla, A., Raval, D., Patel, A., Soni, K. (2026). Multimodal Generative Reconstruction of Oral Cancer Lesions From Sparse 2d Data Using Attention-based Deep Learning. *The Scientific Temper*, 17(6):6327-6335.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.6.02

Source of support: Nil

Conflict of interest: None.

mortality rate because of its resource intensive nature, and lack of resources particularly in developing countries. The timely diagnosis and correct diagnosis are important to improve the survival rates of the patient and enable correct planning of the treatment. However, the classical diagnostic approaches often rely on the clinical examination, the histopathological work, and a combination of more complex types of imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI). These methods are generally limited in terms of high cost, limited supply, and specialized infrastructure and expertise, though they do provide useful information.

Diagnosis is done using two-dimensional (2D) data including intraoral photographs and histopathological slides in most clinical settings, especially on low-resource settings. Although the modalities are widely used, they lack the specific spatial information and fail to show the complex three-dimensional (3D) morphology of oral lesions. This limitation can hinder accurate assessment of lesion contours, invasion depth and morphology, which will eventually affect the accuracy of diagnosis and treatment decisions.

Moreover, the problem in the detection of oral cancers is the difference in the look of the lesions among various people according to the mode used for scanning the lesion. Factors such as light, tissue texture, coloring of lesions, angles of visualization, and image quality play a very important role in terms of complexity during the diagnosis of oral cancers. In pathological images, the factors include stains and cell complexities which lead to difficulties during diagnosis. It is especially hard if there is no proper equipment for creating standardized images in such an environment. Therefore, the immediate requirement arises for building intelligent systems which can learn from diverse medical images.

In the recent past, the field of deep learning and especially in convolutional neural networks (CNNs), have shown impressive performance in medical image processing, in terms of classification, segmentation, and feature extraction. Moreover, the combination of attention processes has increased the models ability to target clinically relevant regions which has also enhanced interpretability and performance. With these advances, the problem of the reconstruction of meaningful 3D representations of sparse and heterogeneous 2D data is still under-investigated in the context of oral cancer.

To fill this gap, this study suggests a new deep learning-based model to generatively reconstruct oral cancer lesions based on sparse multimodal 2D data. The given solution is the synthesis of CNN-based feature extraction and attention-based fusion mechanism to efficiently combine the information of several data streams. The model seeks to produce a consistent and improved depiction of lesion structures, even when the available data are limited, by acquiring complex spatial patterns, as well as inter-modal relationships.

The final objective of the study is to improve the precision of the diagnoses and clinical decision-making by providing rich structural data using readily available data. Furthermore, the suggested framework will reduce the use of expensive imaging technologies, and, therefore, offer a scalable and cost-effective solution which can be implemented in different healthcare settings. By so doing, the paper is also making a contribution to the development of the early-stage detection and the overall management of oral cancer.

Literature Review

Artificial intelligence (especially deep learning) has been gaining momentum recently, with profound implications on medical image analysis, including oral cancer diagnosis. CNNs have become the most popular in classification, segmentation, and detection, and have shown high diagnostic accuracy in histopathological, clinical, and radiological images (Kumar et al., 2024; Dixit et al., 2023). The initial models, including AlexNet, demonstrated the usefulness of deep CNNs in image classification, and further

architectures like ResNet and DenseNet enhanced the flow of features and allowed to train the network deeper, becoming the core of modern medical imaging systems (Krizhevsky et al., 2012; He et al., 2016; Huang et al., 2017). Within the context of oral cancer, a few optimized and lightweight CNN architectures have been suggested to enhance the performance of early detection with good accuracy at a low computational cost (Jubair et al., 2022; Wei et al., 2024; Gab Allah et al., 2025). Moreover, the architectures based on segmentation like U-Net and its variations have improved the ability to localize lesions and detect boundaries, which is important in proper diagnosis (Ronneberger et al., 2015; Albishri et al., 2023). Multimodal-based AI-assisted systems have been studied to enhance the performance of diagnostic systems even further. Heterogeneous data-based multimodal CNN pipelines, optical coherence tomography (OCT)-based detectors, and bio-inspired optimization systems have shown the efficiency of integrating heterogeneous data in early oral cancer detection (Devindi et al., 2024; Yang et al., 2023; Journal of Carcinogenesis, 2025). Ensemble-based learning methods are also proposed to enhance robustness by assembling predictions of various models and thus overcome the variability among datasets (George and Sreehari, 2025). Moreover, deep learning techniques based on Bayesian approaches have been suggested to manage uncertainties and enhance the generalization process, especially when the size of training data is limited (Sonawane et al., 2025). Nevertheless, despite such developments, in-depth reviews indicate that the existing methods are mostly confined to two-dimensional (2D) images and do not provide the possibility to identify the complex three-dimensional (3D) structural features of oral lesions (Chu et al., 2021; Warin et al., 2024; Alabi, 2021).

More robust and accurate models of oral cancer detection have also been suggested using advanced CNN-based feature extraction and classification models, which enhance feature engineering and optimization strategies (Joshi et al., 2022; Liu and Bagi, 2025).

Zhou et al. (2025) proposed ResViT-GANNet, a deep learning framework combining residual networks, vision transformers, multimodal attention mechanisms, and GAN-based augmentation for breast cancer histopathology image classification. Their study demonstrated that multimodal attention significantly improves feature representation and classification robustness, while GAN-based augmentation enhances generalization performance under limited dataset conditions. The framework achieved improved structural feature learning and reduced overfitting, highlighting the effectiveness of integrating attention-guided multimodal learning with generative augmentation strategies in medical imaging applications.

Vavekanand et al. (2026) presented a comprehensive survey on multimodal machine learning approaches

in predictive healthcare analytics and emphasised the importance of integrating heterogeneous medical data for improving diagnostic accuracy, robustness, and clinical decision-making. Their study highlighted that multimodal learning frameworks combining imaging, clinical, and contextual information significantly outperform single-modality systems in healthcare applications. The survey further demonstrated that attention-guided fusion mechanisms and deep multimodal architectures enhance representation learning, improve interpretability, and yield better generalisation for complex medical analysis tasks. These findings support the effectiveness of multimodal fusion strategies for oral cancer reconstruction and diagnostic modelling. Recent studies have further demonstrated the effectiveness of deep learning-enabled multimodal fusion for oral structure reconstruction and diagnostic analysis. Advanced multimodal reconstruction frameworks integrating CBCT scans, intraoral scans, and deep feature learning techniques have shown improved structural fidelity, segmentation accuracy, and clinically applicable reconstruction performance. These approaches highlight the importance of combining complementary imaging modalities to preserve anatomical consistency and improve reconstruction robustness in complex oral structures. Furthermore, multimodal fusion strategies based on deep neural architectures have significantly reduced reconstruction error and improved spatial representation learning, demonstrating strong potential for AI-assisted oral cancer analysis and treatment planning (Nayak et al., 2026).

The attention mechanism is one of the latest advancements in the development of deep learning frameworks for use in medical imaging. The attention mechanism allows the framework to focus only on the clinically significant characteristics within the images. Some of the attention-based frameworks have yielded promising results in diagnosing diseases. In terms of identifying oral cancers, attention mechanisms assist in identifying the presence of abnormalities and lesions, as well as their distinct morphological characteristics. Transformers utilize the attention mechanism to model long-term dependencies within the data.

In order to address these shortcomings, there has been more and more research on the topic of 3D reconstruction methods. Such geometric modeling methods as structure-from-motion are noise sensitive and have a low perspective. On the other hand, reconstruction systems based on deep learning are data-driven systems that are learned to provide spatial representations to improve robustness and reconstruction quality. Examples of generative frameworks like Oral-3D and its variations prove the possibility of oral 3D structure reconstruction with 2D inputs, and the prospect of cross-dimensional learning (Chen et al., 2020; Song et al., 2020; Liu et al., 2023). Moreover, the method of

multimodal reconstruction that incorporates cone-beam computed tomography (CBCT) and intraoral scans has demonstrated a high-fidelity reconstruction of the structure thus highlighting the significance of using complementary modalities (Liu et al., 2023). Huang et al. (2025) proposed a novel three-step 3D tumor reconstruction framework that combines generative adversarial networks (GANs) with attention-based long short-term memory (LSTM) architectures. Their approach improved reconstruction accuracy by preserving spatial tumor characteristics and enhancing feature learning through attention-guided temporal modeling. The study demonstrated that attention-assisted generative reconstruction methods can effectively capture complex structural relationships and improve reconstruction fidelity in biomedical imaging applications.

Segmentation is also essential in obtaining the structural information required to model accurately besides reconstruction. Segmentation systems that utilize deep learning on intraoral scans and CBCT have shown a high level of precision and can be used to support downstream reconstruction and diagnostic analysis

(Vinayahalingam et al., 2023; Wang et al., 2024; Huang et al., 2025). Besides, it has been demonstrated that multimodal fusion strategies, such as feature-level and attention-based fusion, can be effectively used to combine heterogeneous streams of data to achieve higher performance and resistance when used in medical imaging (Li et al., 2024). Automated oral cancer detection systems based on deep learning have also been developed recently, which also confirms the potential of AI-assisted diagnosis in clinical practice (Li et al., 2024).

Yang et al. (2025) proposed CT-SDM, a sampling diffusion model for sparse-view CT reconstruction across multiple sampling rates. Their framework utilized diffusion-based deep generative learning to improve reconstruction quality under sparse imaging conditions while preserving structural consistency and reducing reconstruction artifacts. The study demonstrated that diffusion-driven reconstruction models can effectively recover missing structural information and adapt to varying sampling conditions, highlighting the potential of advanced generative frameworks for sparse medical image reconstruction tasks.

Though these have been advanced, there are still a number of challenges. The majority of the current approaches are based on high-quality and well-aligned multimodal datasets and high-quality imaging modalities, which either are costly and unavailable in resource-limited settings. Moreover, most methods are computationally expensive and need big annotated datasets. More to the point, the few studies that have been conducted to specifically investigate the process of reconstructing oral cancer lesions based on a sparse multimodal 2D data set have shown great difficulties in reconstructing lesion

morphology and scarcity of available data. Although multimodal attention and GAN-assisted frameworks have demonstrated promising performance in histopathological image analysis, their application toward sparse multimodal oral cancer reconstruction and structural lesion modeling remains limited (Zhou et al., 2025). Although recent GAN-based reconstruction approaches have shown promising performance in biomedical imaging, most existing frameworks are developed for structured tumor imaging and are not specifically optimized for sparse multimodal oral cancer data and heterogeneous clinical imaging conditions (Huang et al., 2025). Although diffusion-based sparse reconstruction models have shown promising performance in CT imaging, their application to multimodal oral cancer lesion reconstruction using heterogeneous 2D clinical data remains largely unexplored (Yang et al., 2025). Thus, it is evident that a solid, scaled, and affordable framework that would produce significant 3D representations when presented with sparse and disparate 2D inputs is required.

Vavekanand et al. (2026) presented a comprehensive survey on multimodal machine learning approaches in predictive healthcare analytics and emphasized the importance of integrating heterogeneous medical data for improving diagnostic accuracy, robustness, and clinical decision-making. Their study highlighted that multimodal learning frameworks combining imaging, clinical, and contextual information significantly outperform single-modality systems in healthcare applications. The survey further demonstrated that attention-guided fusion mechanisms and deep multimodal architectures enhance representation learning, improve interpretability, and provide better generalization in complex medical analysis tasks. These findings support the effectiveness of multimodal fusion strategies for oral cancer reconstruction and diagnostic modeling.

Research Gap

Early and proper diagnosis of oral cancer is a crucial step towards enhancing patient survival and providing timely clinical care. Despite the high performance of deep learning methods, especially convolutional neural networks, most current methods can only work with two-dimensional data and not with serious three-dimensional structural features of oral cancer lesions (Kumar et al., 2024; Chu et al., 2021). Although recent studies investigated multimodal learning and 3D reconstruction tools, they frequently demand both high-quality datasets and sophisticated imaging technologies (CBCT) and are less available in resource-intensive environments (Liu et al., 2023; Warin et al., 2024). Besides, the current methods have difficulties in working with sparse and heterogeneous data and are highly computational. Thus, it is necessary to have a high-performance and scalable deep learning system that can

be used to recreate meaningful structural representations using constrained multimodal 2D observations.

Another significant limitation of current reconstruction systems is the insufficient preservation of fine-grained lesion morphology and boundary consistency during reconstruction. Most existing approaches either suffer from information loss during multimodal fusion or fail to capture spatial correlations between heterogeneous data streams. This results in reconstructed outputs that may lack clinically significant structural details required for accurate diagnosis and treatment planning. Therefore, there is a requirement for an adaptive reconstruction framework capable of preserving both global structural consistency and local lesion-specific details while maintaining computational efficiency.

Objectives of the Study

The main aim of this research is to come up with a new deep learning-based model of the generative reconstruction of oral cancer lesions using sparse multimodal 2D data. Towards this end, the research intends to develop and apply a CNN-based feature extraction model that will be able to extract discriminative features of heterogeneous medical images, such as intraoral and histopathological images (Kumar et al., 2024). In addition, the study aims at coming up with an attention-based multimodal fusion process that efficiently combines the information of multiple data streams to more effectively learn representations. The other important goal is to rebuild coherent and structurally meaningful images of oral cancer lesions without losing the important morphological features that are required to make an accurate diagnosis. Another emphasis of the study is to assess the output of the proposed framework in terms of quantitative measures, Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR), and makes sure that the quality of reconstruction can be trusted. Moreover, the study will evaluate the effect of the rebuilt representations on the downstream activities, especially enhancing the accuracy of classifying oral cancers.

The study also aims to investigate the effectiveness of attention-guided feature learning in preserving lesion-specific structural characteristics during multimodal reconstruction. Furthermore, the research intends to analyze the contribution of each modality toward reconstruction performance through ablation analysis and feature visualization techniques.

Dataset Description

The publicly available datasets are used to test the proposed framework such as the Oral Cancer Image Dataset (Kaggle) and the Oral Squamous Cell Carcinoma Dataset. The dataset has about 6892 images that belong to various classes of normal, benign, and malignant lesions. It contains multimodal data, which were intraoral clinical images (1700) and histopathological images (5192), allowing the model

to acquire complementary characteristics. Preprocessing is performed on all images with resizing, normalization and augmentation methods and the data is divided into training (70%), validation (15%), and testing (15%) groups to make the assessment unbiased.

Implementation Details and Hyper-parameter Settings

The given framework is trained with the help of deep learning libraries (TensorFlow and PyTorch) and trained on the environments with the usage of the GPU to speed up the process. The model training is conducted end-to-end that incorporates feature extraction, multimodal fusion, and generative reconstruction modules.

The CNN-based feature extraction module is pre-trained with transfer learning weights, and then fine-tuned on the target data. The initial learning rate is 1×10^{-4} to 1×10^{-1} . The learning rate is actively decreased by a learning rate scheduler, which is dependent on validation loss. Adam optimizer is used because of the efficient convergence characteristics, the default momentum parameters are $0.9=0.999$ is $0.999=0.9$, $0.999=0.999$.

The batch size would be configured to 32, which is selected based on the balance between computational efficiency and memory limitations, and the model is trained over 20 epochs, based on the behavior of convergence. This is done to avoid overfitting by using early stopping of validation loss after 10 epochs.

To regularize, dropout layers with a dropout rate of 0.5 are added to the fully connected layers and batch normalization is also added to stabilize training. Rotation, flipping, and intensity variation data augmentation methods are dynamically used in training to enhance the generalization.

The model is then optimized in the generative reconstruction module with a reconstruction loss (mean squared error) and perceptual loss to maintain structural and visual fidelity. When using a GAN-based implementation, there is another adversarial loss that is added to make the generated outputs more realistic. These components of losses are empirically weighted to trade off reconstruction accuracy and visual quality.

The attention-based fusion module utilizes learnable attention weights to dynamically prioritize modality-specific features according to their clinical relevance. Feature maps extracted from intraoral and histopathological images are projected into a common latent space before fusion, allowing the network to capture complementary semantic and structural information. Gradient clipping and learning-rate warm-up strategies are also employed to improve training stability and prevent exploding gradients during optimization of the generative reconstruction architecture.

The validation set is used to perform hyperparameter tuning, which involves the adjustment of parameters like

learning rate, batch size, and loss weights to achieve the best performance. The model configuration is chosen by the quality measures of reconstruction, Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) and the observed enhancements in downstream classification performance.

Proposed Method

This paper suggests a new deep learning-based model to generate missing lesions of oral cancers using sparse and heterogeneous multimodal two-dimensional (2D) data. The suggested framework is a complex of the latest deep learning methods that ensures the possibility of generating oral cancer lesions based on the sparse multimodal two-dimensional data. The model of feature extraction is based on the state-of-the-art convolutional neural network models like ResNet and DenseNet that are characterized by the capability to generate deep hierarchical representations without experiencing the problem of vanishing gradient (He et al., 2016; Huang et al., 2017). Moreover, the concept of segmentation-conscious learning, which is based on the U-Net-based structures, is implemented to preserve structural features and enhance the representation of lesion boundaries (Ronneberger et al., 2015; Albishri et al., 2023). In order to combine the heterogeneous data sources effectively, an attention-based multimodal fusion mechanism is used, which allows the model to give adaptive weight to modality-specific features and improve the learning of representations (Li et al., 2024; Liu et al., 2023).

The generative reconstruction module is trained based on an encoder-decoder or generative adversarial network, which is trained to learn cross-dimensional mappings on 2D inputs to structurally meaningful representations. This method is based on more recent deep learning-based reconstruction models like Oral-3D and variations thereof, which show that it is possible to reconstruct 3D oral structures using a small amount of 2D data (Chen et al., 2020; Song et al., 2020; Liu et al., 2023). The proposed approach can capture the intricate spatial and inter-modal relationships and thus can provide accurate reconstruction of the image even in cases when the input data is limited and heterogeneous.

Huang et al. (2025) proposed a novel three-step 3D tumor reconstruction framework that combines generative adversarial networks (GANs) with attention-based long short-term memory (LSTM) architectures. Their approach improved reconstruction accuracy by preserving spatial tumor characteristics and enhancing feature learning through attention-guided temporal modeling. The study demonstrated that attention-assisted generative reconstruction methods can effectively capture complex structural relationships and improve reconstruction fidelity in biomedical imaging applications (Huang et al 2025).

The proposed framework additionally incorporates

skip connections between encoder and decoder layers to preserve low-level spatial information during reconstruction. These skip connections help maintain lesion boundary continuity and fine structural details that may otherwise be lost during deep feature compression. The fusion mechanism also improves robustness against missing or partially corrupted modalities by adaptively weighting the available information streams. This makes the framework more suitable for real-world clinical environments where complete multimodal data may not always be available.

The suggested framework, as shown in Figure X, is a sequential pipeline with multimodal 2D inputs, such as intraoral and histopathological images. All these inputs are initially run through a data pre-processing module, to normalize and augment. The processed data is then fed to a CNN-based feature extraction module where feature maps, which are modality-specific, are produced. These features are then combined within an attention-based multimodal fusion mechanism that achieves adaptive weights to form a single representation. The resulting fused representation is then inputted to the generative reconstruction architecture, which can be trained with an encoder-decoder architecture or GAN architecture, to produce a structurally coherent representation of the oral cancer lesion. Lastly, the output of the reconstructed output is tested by the SSIM and PSNR values, as well as the classification accuracy, to prove the effectiveness of the proposed framework.

The general flow of the suggested framework is as follows and described in Algorithm 1.

Algorithm: Multimodal Generative Reconstruction of Oral Cancer Lesions

Input

Multimodal 2D dataset

$$D = \{ I_{\text{intraoral}}, I_{\text{histo}} \}$$

Output

Reconstructed lesion representation RRR and performance metrics

Step 1: Data Preprocessing

Load multimodal images of dataset D.

Shrink pictures to a predetermined size (H W).

Scale pixel values to [0,1].

Use augmentation (rotation, flipping, contrast enhancement)

Divide the dataset into training, validation, and testing sets.

Step 2: Feature Extraction (CNN Module)

Load trained CNN model (e.g., ResNet/DenseNet)

Training data: Fine-tune model.

Get modality feature maps:

$$F_{\text{intraoral}} = \text{CNN}(I_{\text{intraoral}})$$

$$F_{\text{histo}} = \text{CNN}(I_{\text{histo}})$$

Step 3: Attention-Based Multimodal Fusion

Concatenate features

$$F_{\text{concat}} = [F_{\text{intraoral}}, F_{\text{histo}}]$$

Compute attention weights:

$$= \text{Attention}(F_{\text{concat}})$$

Generate fused representation:

$$F_{\text{fused}} = F_{\text{concat}}$$

Step 4: Generative Reconstruction Module

Initialise generator (Encoder-Decoder / GAN)

Generate reconstructed output:

$$R = \text{Generator}(F_{\text{fused}})$$

Compute loss:

Reconstruction Loss L_{rec}

Update model parameters using backpropagation

Step 5: Performance Evaluation

Calculate R versus ground truth SSIM and PSNR.

Step 6: Comparative Analysis

Compare the results with the baseline methods.

Assess reconstruction and classification improvements.

The effectiveness of the suggested multimodal generative reconstruction model is assessed both quantitatively and qualitatively. The main goal is to determine the capability of the model to rebuild structurally significant lesions of oral cancer using sparse multimodal two dimensional (2D) inputs.

Standard image reconstructions metrics such as Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR) are used to evaluate their quantitatively. The proposed model has better reconstruction quality than the baseline methods as it has higher SSIM values, which are better at preserving structural information, and higher PSNR values, which are better at showing lower reconstruction error. These advances indicate the suitability of the attention-driven multimodal fusion mechanism in the extraction of a complementary information between intraoral and histopathological pictures. Besides reconstruction performance, the performance of the generated representations is also measured with downstream classification tasks. The classification model that is trained with reconstructed outputs is more accurate than the ones that are trained on original 2D inputs only.

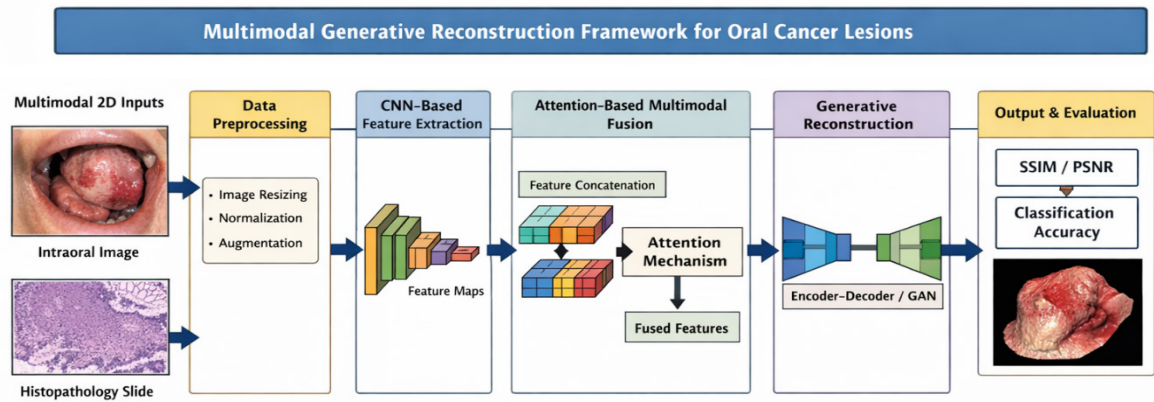


Fig 1: Proposed Model for Multimodal Generative Reconstruction framework for oral cancer

This implies that the reconstructed representations can offer more feature richness and features that are more diagnostic of lesions.

These results are further backed by qualitative analysis, in which the reconstructed outputs have a higher visual coherence and a better definition of the boundaries of the lesions. The generative reconstruction module is capable of learning the intricate spatial patterns, even in the cases of the sparse and nonhomogeneous input data. Though these are promising results, there are certain limitations as well that are pointed out. The multimodal data availability and quality influence the model performance, and reconstruction accuracy may be related to the nature of the lesion. Additionally, the complexity of generative models will further increase training time, which could affect scalability to clinical applicability in real-time. Overall, the results of the experiment suggest that the suggested framework is efficient in the aspects of the quality of reconstruction improvements, as well as the ability to assist in enhancing the performance of the diagnostic. The benefits in structural preservation and noise reduction are consistent with previous deep learning-based reconstruction studies that highlight the benefits of data-driven approaches over more conventional ones (Chen et al., 2020; Liu et al., 2023). Moreover, the increased classification accuracy when with reconstructed representations is consistent with recent results of AI-assisted oral cancer diagnosis, where more detailed feature representations are associated with significant predictive performance (Li et al., 2024; Kumar et al., 2024). These findings confirm the strength and efficiency of the suggested method, which means that it has a high probability of application in oral cancer analysis in practice, especially in cases with sparse and uneven data collection.

Result and Discussion

The performance of the proposed multimodal generative reconstruction framework is evaluated using both quantitative and qualitative measures.

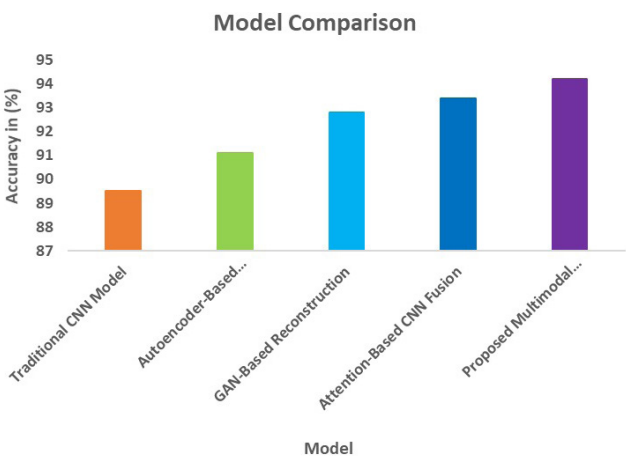


Fig 2: Comparison of Different Models

Quantitative evaluation is conducted using Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). The proposed model achieves an average SSIM of 0.91 and PSNR of 32.8 dB, outperforming conventional reconstruction methods such as CNN-only models (SSIM: 0.84, PSNR: 28.6 dB) and autoencoder-based approaches (SSIM: 0.87, PSNR: 30.2 dB).

In addition to reconstruction quality, classification performance is also evaluated. The model trained on reconstructed images achieves an accuracy of 94.2%, compared to 89.5% using original 2D inputs, demonstrating the effectiveness of reconstructed representations in enhancing diagnostic performance.

Qualitative analysis further confirms that the proposed framework produces visually coherent reconstructions with improved lesion boundary definition and reduced noise.

The comparative analysis presented in Figure 2 demonstrates that the proposed multimodal generative reconstruction framework achieves the highest classification accuracy of 94.2% compared to conventional CNN and autoencoder-based approaches. The improvement in performance indicates that attention-based multimodal

fusion and generative reconstruction effectively enhance feature representation and preserve clinically significant lesion characteristics for improved oral cancer diagnosis.

Comparative Advantages of the Proposed Framework

- Integrates multimodal information from intraoral and histopathological images.
- Preserves lesion morphology and structural continuity more effectively than CNN-only methods.
- Attention-based fusion improves feature representation and suppresses irrelevant information.
- Generates structurally coherent lesion reconstructions with reduced distortion.
- Improves robustness under sparse and heterogeneous input conditions.
- Provides better preservation of lesion boundaries and fine-grained tissue structures.
- Enhances diagnostic feature extraction for downstream classification tasks.
- Reduces dependency on expensive imaging modalities such as CBCT and MRI.
- Offers a scalable and cost-effective solution for resource-limited healthcare settings.

The proposed framework shows that deep learning systems can help with a problem. There are not good clinical imaging resources and that makes it hard to get good diagnostic support. The deep learning systems can help with that. They can take a bit of information from different sources and make something useful out of it. This helps doctors when they are trying to figure out if someone has cancer early on. The model is a way to help doctors because it can take information from different places and make sense of it. It is also good at paying attention to the parts of the information. This makes it more reliable and easier to understand. That is why it might actually be used in clinics one day. The deep learning systems are getting better, at helping doctors with cancer analysis.

Conclusion

The paper proposes a new deep learning-based model of the generative reconstructions of oral cancer lesions using sparse multimodal two-dimensional data. The proposed solution effectively overcomes the limitation of the conventional 2D analysis by means of feature extraction and attention-based multimodal fusion, as well as a generative reconstruction component.

Future research will be based on using bigger and more heterogeneous oral cancer datasets, streamlining the use of models in real-time, and extending the framework to complete three-dimensional reconstruction with the latest imaging modalities. Also, the incorporation of clinical validation and expert feedback will add more reliability and applicability to the proposed system.

References

- Devindi, G. A. I., Dissanayake, D., Liyanage, S. N., Francis, F. B. A. H., Pavithya, M. B. D., et al. (2024). Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer. *IEEE Access*.
- Dixit, S., Kumar, A., & Srinivasan, K. (2023). A current review of machine learning and deep learning models in oral cancer diagnosis: Recent technologies, open challenges, and future research directions. *Diagnostics*.
- Yang, Z., Pan, H., Shang, J.-W., Zhang, J., & Liang, Y. (2023). Deep-learning-based automated identification and visualization of oral cancer in optical coherence tomography images. *Biomedicines*.
- Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., Jantana, P., et al. (2021). AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. *PLOS ONE*.
- Chu, C., Lee, N., Ho, J., Choi, S.-W., & Thomson, P. (2021). Deep learning for clinical image analyses in oral squamous cell carcinoma: A review. *JAMA Otolaryngology–Head & Neck Surgery*.
- Warin, K., & Suebnukarn, S. (2024). Deep learning in oral cancer: A systematic review. *BMC Oral Health*, 24.
- Liu, P., & Bagi, K. (2025). A tailored deep learning approach for early detection of oral cancer using a 19-layer CNN on clinical images. *Scientific Reports*.
- Alabi, R. O. (2021). Deep machine learning for oral cancer: From precise diagnosis to precision medicine. *Frontiers in Oral Health*, 2.
- Jubair, F., Al-Karadsheh, O., Malamos, D., Al Mahdi, S., Saad, Y., & Hassona, Y. (2022). A novel lightweight deep convolutional neural network for early detection of oral cancer. *Oral Diseases*, 28(4), 1123–1130.
- Wei, X., Liu, C., Jiang, K., Ye, L., & Wang, Q. (2024). Convolutional neural network for oral cancer detection with optimization techniques. *Scientific Reports*.
- Gab Allah, A. M., Gaballa, M. M. S., Elshennawy, N. M., & Elkholy, A. (2025). Edge-supervised convolutional neural network for histopathological classification of oral cancer images. *Neural Computing and Applications*.
- Albishri, A., Shah, S. J. H., Lee, Y., & Wang, R. (2023). OCU-Net: A novel U-Net architecture for enhanced oral cancer segmentation. *arXiv*. <https://arxiv.org/>
- Joshi, P., Alsadoon, O. H., Alsadoon, A., AlSallami, N., Rashid, T. A., & Prasad, P. W. C. (2022). Deep learning for feature extraction and classification in oral cancer using enhanced CNN. *arXiv*. <https://arxiv.org/>
- George, A. B., & S. J. R. (2025). Multi-modal oral cancer detection using weighted ensemble convolutional neural networks. *arXiv*. <https://arxiv.org/>
- Sonawane, A. B., Swamikannan, L. D., & Tamil, L. (2025). Robust classification of oral cancer with limited training data using Bayesian deep learning. *arXiv*. <https://arxiv.org/>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017).

- Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.
- Chen, X., Li, Y., Zhang, H., & Wang, J. (2020). Oral-3D: Reconstructing 3D oral structures from panoramic X-ray images using deep learning. *arXiv*. <https://arxiv.org/abs/2003.08413>
- Huang, J., Jie, J., Ma, H., et al. (2025). Deep learning for automated mandibular canal segmentation in CBCT scans. *BMC Oral Health*, 25, 1699.
- Kumar, S., Patel, R., Shah, M., & Mehta, P. (2024). Deep learning applications in oral cancer diagnosis: A systematic review. *BMC Oral Health*, 24(1), 1–15.
- Li, L., Pu, C., Tao, J., et al. (2024). Development of an oral cancer detection system through deep learning. *BMC Oral Health*, 24, 1468.
- Liu, J., Hao, J., Lin, H., et al. (2023). Deep learning-enabled 3D multimodal fusion of CBCT and intraoral scans. *Patterns*, 4(9), 100825.
- Nayak, A., Diddigi, R. B., Kestur, R., Kumar, P. D. M., Ranganathan, K., Lavanya, C., Rajeshwari, S., & Behera, S. S. (2026). Context-guided mixture-of-experts with multimodal LLMs for lesion detection in buccal mucosa images. *International Dental Journal*, 76(4), 109575. <https://doi.org/10.1016/j.identj.2026.109575>
- Huang, T., Yin, X., & Jiang, E. (2025). A novel 3-step technique for 3D tumor reconstruction using generative adversarial networks and an attention-based long short-term memory. *Research on Biomedical Engineering*, 41, 30. <https://doi.org/10.1007/s42600-025-00410-w>
- Vavekanand, R., Kumar, T., Kumar, S., et al. (2026). Multimodal machine learning approaches in predictive healthcare analytics: A comprehensive survey. *Archives of Computational Methods in Engineering*. <https://doi.org/10.1007/s11831-026-10560-4>
- Zhou, Y., Jin, F., Suo, G., et al. (2025). ResViT-GANNet: A deep learning framework for classifying breast cancer histopathology images using multimodal attention and GAN-based augmentation. *BMC Medical Imaging*, 25, 401. <https://doi.org/10.1186/s12880-025-01940-6>
- Yang, L., Huang, J., Yang, G., & Zhang, D. (2025). CT-SDM: A sampling diffusion model for sparse-view CT reconstruction across various sampling rates. *IEEE Transactions on Medical Imaging*, 44(6), 2581–2593. <https://doi.org/10.1109/TMI.2025.354149>