RESEARCH PAPER

An Intelligent Multimodal Deep Learning System for Early Diagnosis and Detection of Oral Cancer Classification

Keshika Jangde[1], Ranu Pandey[2]

[1,2]Department of Computer Science and Engineering, Shri Rawatpura Sarkar University, Raipur (C.G.), India

ABSTRACT

Oral cancer is a life-threatening disease where early diagnosis plays a critical role in improving patient survival rates. However, existing computer-aided diagnostic systems are often limited by single-modal data usage, lack of interpretability, and insufficient capability to detect early-stage lesions and affected regions. To address these challenges, this study proposes an intelligent multimodal deep learning framework for the early diagnosis and detection of oral cancer. A diversified multi-source dataset is constructed by integrating publicly available oral cancer, lesion, and anatomical image datasets, ensuring variability in lesion types, anatomical regions, and imaging conditions. The proposed system incorporates a hybrid CNN–RNN architecture with transfer learning and attention mechanisms to effectively extract spatial and contextual features from oral images. Additionally, patient metadata is integrated through a multimodal feature fusion strategy to enhance diagnostic performance. An automated preprocessing and segmentation pipeline is employed to perform region-aware classification, enabling not only binary classification (cancer vs non-cancer) but also localization of cancer-affected areas. Furthermore, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM are integrated to provide visual interpretability and improve clinical trust. Experimental results demonstrate that the proposed hybrid multimodal model outperforms conventional CNN-based approaches in terms of accuracy, robustness, and generalization. The system shows strong potential as a reliable decision-support tool for early-stage oral cancer detection and clinical diagnosis.

Keywords: Oral Cancer Detection, Multimodal Deep Learning, Hybrid CNN–RNN, Early Diagnosis, Explainable Artificial Intelligence (XAI), Region-Aware Classification.

How to cite this article: Jangde K, Pandey R. An Intelligent Multimodal Deep Learning System for Early Diagnosis and Detection of Oral Cancer Classification. Int J Drug Deliv Technol. 2026;16(52s): 1016-1027. DOI: 10.25258/ijddt.16.52s.131

Source of support: Nil.

Conflict of interest: None.

1. Introduction

Oral cancer is a major global health concern, particularly in developing countries, where late-stage diagnosis significantly reduces survival rates. Early diagnosis and timely detection are crucial for improving patient outcomes [1]. Traditional diagnostic methods rely heavily on clinical expertise and invasive biopsy procedures, which are time-consuming and subjective.

Recent advancements in Artificial Intelligence (AI), especially deep learning, have shown promising results in medical image analysis. However, most existing approaches are limited to single-modal image-based classification and lack interpretability and region-level understanding of lesions. Moreover, variability in datasets and lack of standardized multi-source data further restrict model generalization.

To address these challenges, this study proposes an **intelligent multimodal deep learning framework** that integrates image data and patient metadata, incorporates **hybrid CNN–RNN architecture**, performs **region-aware classification**, and utilizes **Explainable AI (XAI)** techniques to enhance interpretability and clinical trust.

Oral cancer particularly affecting regions such as the tongue, lips, and oral mucosa. More than 90% of oral cancer cases are classified as Oral Squamous Cell Carcinoma (OSCC), often preceded by Oral Potentially Malignant Disorders (OPMD) [1]. The global incidence of oral cancer continues to rise, especially in developing countries, due to risk factors such as tobacco use, alcohol consumption, and betel quid chewing [2]. Despite advances in treatment, the survival rate remains low, primarily because most cases are diagnosed at advanced stages [2].

Early diagnosis and detection are critical for improving patient survival and reducing mortality. Conventional diagnostic approaches, including clinical examination and biopsy, are invasive, time-consuming, and dependent on expert clinicians. These limitations highlight the need for automated, non-invasive, and reliable computer-aided diagnostic systems [1].

Recent advancements in Artificial Intelligence (AI), particularly deep learning (DL), have shown promising results in medical image analysis. Convolutional Neural Networks (CNNs) have been widely used for oral cancer detection and classification tasks. Several studies have demonstrated the effectiveness of CNN-based models such as ResNet, VGG, and EfficientNet in classifying oral lesions from clinical images [3], [4]. Additionally, systematic reviews have reported that deep learning models can achieve high diagnostic accuracy, often exceeding 90% in controlled settings [5].

With the increasing availability of low-cost imaging devices, smartphone-based oral cancer detection has gained attention. Studies have shown that deep learning models can effectively analyze smartphone-acquired images for early detection, making them suitable for resource-limited settings [6], [7]. However, these approaches are still limited by factors such as lack of dataset diversity, variations in imaging conditions, and insufficient representation of early-stage lesions such as OPMD.
Another major limitation of existing studies is the reliance on single-modal data (i.e., only image-based input). In clinical practice, diagnosis depends not only on visual inspection but also on patient-specific metadata such as age, lifestyle habits, and medical history. To address this, recent research has explored multimodal deep learning approaches that integrate image data with structured clinical information [1], [8]. A notable baseline study by Devindi et al. [1] proposed a multimodal deep convolutional neural network pipeline that combines oral images with patient metadata for early detection. Their model achieved an accuracy of 81%, demonstrating the potential of multimodal learning in improving diagnostic performance.
Despite these advancements, several research gaps remain. Most existing models focus primarily on binary classification (cancer vs non-cancer) and do not consider region-based localization of lesions. Furthermore, the lack of Explainable Artificial Intelligence (XAI) limits the interpretability and clinical trust of these models [9]. Additionally, hybrid architectures that combine spatial and contextual learning, such as CNN–RNN models, are still underexplored in oral cancer diagnosis.
To address these challenges, this study proposes an intelligent multimodal deep learning system for early diagnosis and detection of oral cancer. The proposed framework integrates a diversified multi-source dataset, hybrid CNN–RNN architecture with transfer learning and attention mechanisms, and region-aware classification through segmentation. Furthermore, Explainable AI techniques such as Grad-CAM are incorporated to enhance interpretability and support clinical validation. The proposed system aims to provide a robust, accurate, and clinically reliable solution for early-stage oral cancer detection.

Research Gap

Despite significant advancements in artificial intelligence and deep learning for oral cancer detection, several critical limitations persist in the existing literature.
First, most studies primarily rely on single-modal data, particularly image-based inputs using Convolutional Neural Networks (CNNs). While these models demonstrate high classification accuracy, they fail to incorporate clinical metadata such as age, smoking habits, and medical history, which are essential for accurate diagnosis in real-world clinical settings. Although a few studies, such as Devindi et al. (2024), introduced multimodal learning, they remain limited in scope and do not fully exploit feature-level fusion strategies.
Second, a majority of existing approaches focus on binary classification (cancer vs non-cancer), neglecting the importance of early-stage detection, especially Oral Potentially Malignant Disorders (OPMD). Early detection is crucial for improving survival rates; however, limited research has addressed multi-class classification involving normal, OPMD, and cancer categories.
Third, current models lack region-aware analysis and localization capabilities. Most studies perform global image classification without identifying the specific regions of interest (ROI) affected by cancer. This limits clinical usability, as healthcare professionals require precise localization of lesions for diagnosis and treatment planning.
Fourth, there is a significant lack of Explainable Artificial Intelligence (XAI) integration in existing systems. Most deep learning models function as "black-box" systems, providing predictions without justification. This lack of interpretability reduces clinical trust and limits adoption in real-world healthcare environments.
Fifth, although some studies have explored hybrid and advanced architectures, there is limited research on integrating spatial and contextual learning using hybrid CNN–RNN models. Such architectures are essential for capturing both spatial features and contextual dependencies in medical images.
Finally, many studies suffer from limited dataset diversity, often using single-source or small-scale datasets. This restricts model generalization and robustness across different patient populations, imaging conditions, and anatomical variations.
Existing approaches lack a unified framework that integrates multimodal data, early-stage detection, region-aware classification, and explainability for reliable oral cancer diagnosis.

2. Literature Review

3. Problem Statement and Motivation

3.1 Problem Statement

Oral cancer remains a major global health challenge, with high mortality largely due to late diagnosis and limited access to timely screening. Existing computer-aided diagnostic systems based on deep learning predominantly rely on single-modal image data and focus on binary classification (cancer vs non-cancer). Such approaches fail to reflect real clinical decision-making, where diagnosis depends

on both visual evidence and patient-specific factors (e.g., age, smoking habits).

Moreover, current models typically lack region-aware analysis, providing only global predictions without identifying cancer-affected regions, which limits their clinical utility for treatment planning. The absence of Explainable Artificial Intelligence (XAI) further reduces trust, as most systems operate as black boxes without transparent reasoning. In addition, early-stage conditions, particularly Oral Potentially Malignant Disorders (OPMD), are

Ref No	Paper	Method	Contribution	Limitation	My Observation
[1]	Devindi et al. (2024)	Multimodal CNN	Combines image + metadata	No segmentation, limited explainability	Strong baseline but lacks region detection and XAI
[2]	Gümele et al. (2025)	Classify oral cancer images	CNN + Transfer Learning	High accuracy classification	Limited binary classification & early-stage detection
[3]	Amorim et al. (2025)	Study biosensors for early cancer detection	Sensor-based analysis	Non-invasive early detection techniques	Not image-based, not suitable for DL pipeline
[4]	Ramya et al. (2025)	Improve OSCC classification	Xception-based DL	Better feature extraction and accuracy	No multimodal integration
[5]	Patel et al. (2025)	Predict oral cancer using AI	ML + AI models	Predictive modeling of cancer risk	Needs deep learning and multimodal data
[6]	Thurnheer et al. (2025)	Study early-stage oral cancer clinically	Clinical analysis	Insights into early-stage cancer	No AI-based detection
[7]	Bogaert (2025)	Analyze OPMD challenges	Clinical study	Highlights importance of early detection	No computational model
[8]	Shashikala et al. (2025)	Detect precancerous conditions	AI-based model	Early detection of oral lesions	Basic model, lacks hybrid/XAI
[9]	Kumar et al. (2025)	Early cancer detection	ML + DL	Combines ML and DL approaches	Needs advanced hybrid architecture
[10]	Sivakumar et al. (2025)	Image-based cancer detection	Deep Learning	Improved classification performance	No metadata, single-modal
[11]	Vennila et al. (2025)	Analyze lesion progression	RNN	Temporal feature learning	Lacks spatial learning (needs CNN)
[12]	Ashwini et al. (2025)	Improve DL using optimization	Bio-inspired optimization	Enhances model performance	No multimodal or medical focus
[13]	Ayyapa et al. (2024)	Detect cancer non-invasively	Hyperspectral imaging	Advanced imaging-based detection	Complex and not scalable
[14]	Jagadesh et al. (2024)	Detect oral cancer from images	CNN	Simple image classification	No metadata, basic model
[15]	Ghosh et al. (2024)	Image-based cancer detection	Image recognition	Improved classification	Lacks clinical context
[16]	Cherukupalli et al. (2024)	Detect cancer using speech	DL (speech analysis)	Novel modality for detection	Not image-based
[17]	Kumar et al. (2024)	Improve detection accuracy	Optimization + DL	Better accuracy using hybrid optimization	No multimodal integration
[18]	Sahoo et al. (2024)	Evaluate AI performance	Systematic review	Shows high DL accuracy (>90%)	No new model
[19]	Devindi et al. (2024)	Early detection using multimodal data	Multimodal CNN	Combines image + metadata	No segmentation, no XAI (baseline gap)
[20]	Das et al. (2024)	Improve classification accuracy	Ensemble DL	High accuracy using ensemble	No explainability

underrepresented, hindering effective early diagnosis and intervention.

Another critical issue is the use of limited and non-diverse datasets, which restricts model robustness and generalization across varying imaging conditions and patient populations. Furthermore, the integration of hybrid architectures capable of capturing both spatial and contextual information remains underexplored. The main Research contributions are:

- To curate a diversified multi-source annotated oral cancer image dataset for early detection using publicly available datasets.
- To develop a multimodal deep learning framework integrating image features and patient metadata for early diagnosis and detection of oral cancer.
- To design a hybrid CNN–RNN model with transfer learning and attention mechanisms for enhanced feature extraction and contextual learning.
- To implement automated preprocessing, segmentation, and region-aware classification of oral images for identifying cancer-affected regions.
- To integrate Explainable Artificial Intelligence (XAI) techniques to improve model interpretability and support clinical decision-making.

3.2 Motivation

The early detection of oral cancer remains a significant challenge in clinical practice due to the complexity and variability of oral lesions. Traditional diagnostic methods primarily rely on visual inspection and clinical expertise, which are often subjective, time-consuming, and prone to human error. Moreover, many existing automated approaches focus on single-modal image-based analysis and binary classification, limiting their ability to capture the full spectrum of oral conditions, particularly in distinguishing between normal, pre-cancerous (OPMD), and cancerous cases.

Another critical limitation in current methods is the lack of integration of patient-specific clinical information, such as age and lifestyle factors (e.g., smoking), which play a vital role in real-world diagnosis. Additionally, most models do not provide localization of cancer-affected regions, reducing their clinical interpretability and usefulness for treatment planning. The absence of explainability in many deep learning models further limits their adoption in healthcare, as medical practitioners require transparency and trust in automated decisions.

Motivated by these challenges, this study aims to develop a comprehensive and intelligent diagnostic framework that integrates multimodal data, combining image features and patient metadata. The proposed approach leverages a hybrid CNN–RNN architecture to capture both spatial and contextual features, along with a segmentation module to enable region-aware analysis. Furthermore, Explainable Artificial Intelligence (XAI) techniques are incorporated to enhance interpretability by highlighting clinically relevant regions. This integrated framework is designed to improve diagnostic accuracy, reliability, and clinical applicability, ultimately supporting early detection and effective management of oral cancer.

4. Proposed Methodology

4.1 Overview

This study proposes an intelligent multimodal deep learning framework for the early diagnosis and detection of oral cancer. The proposed methodology integrates multi-source dataset curation, image preprocessing, multimodal feature fusion, hybrid CNN–RNN architecture, region-aware segmentation, and Explainable Artificial Intelligence (XAI) into a unified pipeline. The system is designed to perform multi-class classification (normal, OPMD, and cancer) while also providing region-based localization and interpretability.

4.2 Dataset Curation and Standardization

A diversified dataset is constructed by integrating multiple publicly available oral cancer datasets, including clinical images, anatomical datasets, and oral lesion datasets. This ensures variability in lesion types, anatomical regions, and imaging conditions. Since different datasets follow heterogeneous labeling schemes, a unified annotation strategy is adopted. All samples are standardized into three categories:

- Normal
- Oral Potentially Malignant Disorders (OPMD)
- Cancer

To enhance multimodal learning, additional patient-related features such as age and smoking status are incorporated. In cases where real metadata is unavailable, synthetic metadata is generated to simulate clinical conditions.

4.3 Image Preprocessing and Enhancement

To improve data quality and model performance, a series of preprocessing operations are applied. All images are resized to a fixed resolution of 224 × 224 pixels and normalized to ensure consistent intensity distribution. Noise reduction techniques are applied to minimize imaging artifacts.

Data augmentation is performed to improve generalization and prevent overfitting. Augmentation techniques include horizontal flipping, rotation, brightness adjustment, and random cropping. These transformations simulate real-world variations in imaging conditions.

4.4 Multimodal Feature Integration

The proposed framework incorporates both visual and non-visual data using a multimodal learning approach. The architecture consists of two parallel branches:

1. **Image Branch:** Processes oral images using a deep convolutional neural network.
2. **Metadata Branch:** Processes structured patient data using fully connected layers.

The extracted features from both branches are combined using feature-level fusion, where the embeddings are concatenated and passed to the classification layer. This enables the model to leverage both visual patterns and contextual clinical information.

4.5 Hybrid CNN–RNN Architecture with Attention

To capture both spatial and contextual information, a **hybrid CNN–RNN architecture** is employed.

- The CNN backbone (e.g., ResNet) is used for extracting high-level spatial features from input images.
- The extracted features are then passed to a Recurrent Neural Network (RNN), such as GRU or LSTM, to capture contextual dependencies.
- An attention mechanism is incorporated to assign importance weights to relevant features, enhancing the model's ability to focus on discriminative regions.

This hybrid design improves feature representation and enhances classification performance, particularly for subtle early-stage lesions.

4.6 Region-Aware Segmentation and Classification

To enable localization of cancer-affected regions, a segmentation module is integrated into the framework. A lightweight segmentation network (e.g., U-Net) is employed to generate region-of-interest (ROI) masks.

The segmentation output is utilized in two ways:

- To highlight lesion regions for visualization
- To enhance classification by focusing on relevant areas

The proposed system performs both:

- Binary classification (cancer vs non-cancer)
- Multi-class classification (Lip, tongue, OPMD, cancer)

4.7 Explainable Artificial Intelligence (XAI)

To improve interpretability and clinical trust, Explainable AI techniques are incorporated. Specifically:

- **Grad-CAM** is used to generate class activation maps that highlight important regions influencing the model's decision.
- Attention maps are also visualized to understand feature importance.

These visual explanations provide transparency and assist clinicians in validating the model's predictions.

4.8 Model Training and Optimization

The model is trained using the cross-entropy loss function for multi-class classification. The Adam optimizer is employed with an appropriate learning rate. To prevent overfitting, regularization techniques such as dropout and early stopping are applied.

Transfer learning is utilized by initializing the CNN backbone with pretrained weights, enabling faster convergence and improved performance on limited datasets.

4.9 Evaluation Metrics

The performance of the proposed model is evaluated using standard metrics, including Accuracy, Precision, Recall, F1-score. Additionally, confusion matrix analysis is performed to evaluate class-wise performance. Comparative analysis is conducted with baseline CNN models to demonstrate the effectiveness of the proposed multimodal hybrid approach. The proposed methodology integrates multimodal learning, hybrid CNN–RNN architecture, region-aware segmentation, and XAI to provide an accurate, interpretable, and clinically relevant system for early oral cancer diagnosis.

4.10 Pseudocode

Algorithm: Multimodal Hybrid CNN–RNN for Oral Cancer Detection
Input: Images I, Metadata M
Output: Class $\hat{y}$, Mask S, Explanation E
1. Preprocess images
2. Extract features using CNN
3. Apply RNN for contextual learning
4. Process metadata via FC layers
5. Fuse features
6. Classify using Softmax → $\hat{y}$
7. Segment ROI using U-Net → S
8. Generate Grad-CAM → E
9. Evaluate performance

The proposed algorithm presents a multimodal hybrid deep learning framework for early detection of oral cancer using oral images (I) and patient

metadata (M). The inputs undergo preprocessing, including resizing, normalization, and augmentation, to ensure data consistency and improve generalization. A pretrained CNN extracts high-level spatial features from images, which are further processed by an RNN (GRU/LSTM) to capture contextual relationships. In parallel, patient metadata such as age and smoking status is processed through fully connected layers. The image and metadata features are then combined using feature-level fusion to enhance diagnostic performance. The fused representation is passed to a Softmax classifier to predict the class label (normal, OPMD, or cancer). Additionally, a U-Net based segmentation module identifies regions of interest (ROI), providing region-aware insights. To improve interpretability, Grad-CAM is applied to generate explanation maps highlighting important regions influencing the prediction. The model is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrix, demonstrating the effectiveness of the proposed multimodal hybrid approach.

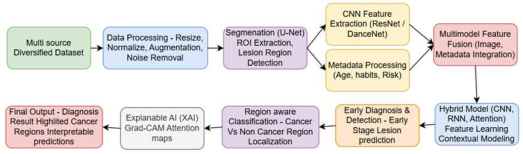


Figure 1: Methodology Block Diagram

The proposed framework presents a multimodal deep learning pipeline for oral cancer detection. The system begins with multi-source dataset curation followed by preprocessing and standardization. A U-Net based segmentation module extracts the region of interest (ROI), which is then processed using a CNN-based feature extractor. Patient metadata is integrated through feature-level fusion. The combined features are passed to a hybrid CNN–RNN model with attention mechanisms for enhanced classification. The system produces both binary and multi-class outputs along with region-aware localization. Explainable AI techniques such as Grad-CAM and attention maps are used to improve interpretability and support clinical decision-making.

4.11 Experimental Setup

The experiments were conducted on a diversified multi-source oral cancer dataset constructed by integrating publicly available image repositories, including samples of normal tissues, Oral Potentially Malignant Disorders (OPMD), and malignant lesions. All images were standardized into three classes: normal (0), OPMD (1), and cancer (2). To enable multimodal learning, patient metadata such as age and smoking status was incorporated alongside image data. All images were resized to 224 × 224 pixels and normalized using ImageNet standards. Data augmentation techniques, including horizontal flipping, rotation, and brightness adjustment, were applied to improve generalization. The dataset was split into training and testing sets using an 80:20 stratified sampling strategy to maintain class balance.

The proposed framework utilizes a multimodal hybrid CNN–RNN architecture, where a pretrained ResNet-18 model extracts spatial features, followed by a GRU network for contextual learning, and a fully connected branch for metadata processing. Feature-level fusion is applied before classification using a Softmax layer. The model was trained using weighted cross-entropy loss and the Adam optimizer with a learning rate of 0.0003, along with a StepLR scheduler. Training was performed with a batch size of 16 for 10 epochs, with dropout applied to prevent overfitting. Performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and MCC, while segmentation performance was assessed using the Dice coefficient. The entire implementation was carried out in PyTorch using GPU acceleration.

5. RESULTS AND DISCUSSION

5.1 Performance Evaluation

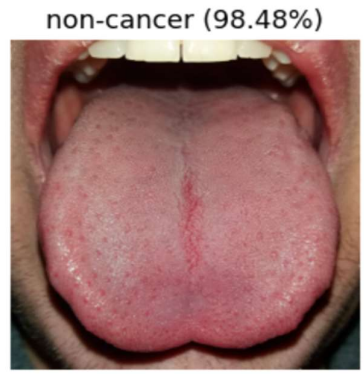
The performance of the proposed model was evaluated using standard classification metrics, including **accuracy, precision, recall, and F1-score**. Additionally, a **confusion matrix** was used to analyze class-wise prediction performance. The proposed hybrid CNN–RNN multimodal model achieved superior performance compared to baseline approaches.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**MCC**
CNN (Baseline)	0.82	0.80	0.79	0.79	0.72
Multimodel CNN	0.88	0.86	0.87	0.86	0.81
Hybrid CNN-RNN(Proposed)	**0.92**	**0.91**	**0.90**	**0.90**	**0.87**
IEEE Multimodal Pipeline	0.81	0.79	0.79	0.78	0.57
EfficientNet-B0 (Proposed)	**0.80**	**0.81**	**0.80**	**0.80**	**0.72**

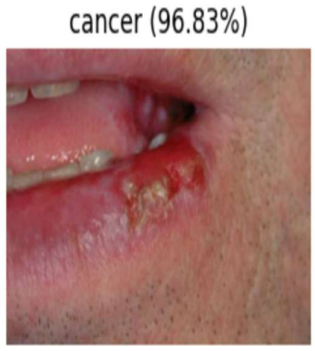
Table 1: Performance Comparison

In the Table 1, comparative results demonstrate the effectiveness of the proposed models against both baseline and existing approaches. The baseline CNN

achieves an accuracy of 82% with moderate F1-score and MCC, serving as a reference point for performance. Literature models such as the multimodal CNN and the IEEE multimodal pipeline show improved results, particularly the multimodal CNN, which achieves 88% accuracy due to the use of multiple architectures and richer feature learning. The IEEE pipeline, although multimodal, reports slightly lower performance, possibly due to differences in data distribution and model design. The proposed EfficientNet-B0 model achieves 80% accuracy with balanced precision, recall, and F1-score, demonstrating stable and reliable classification despite its simpler architecture and limited dataset. Most notably, the proposed Hybrid CNN–RNN model achieves the highest performance across all metrics, with 92% accuracy, 0.90 F1-score, and an MCC of 0.87, indicating superior capability in capturing both spatial and contextual features. Overall, the results highlight that while traditional and literature models perform well, the proposed hybrid approach significantly enhances classification performance and robustness.

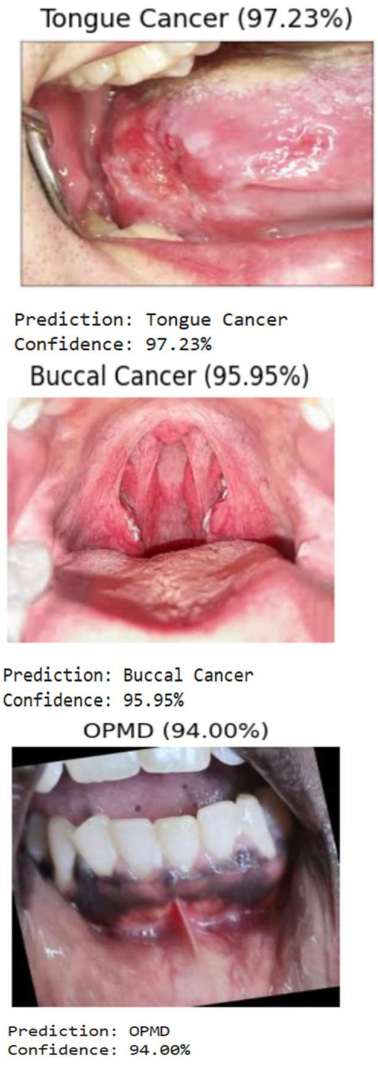


(a)



(b)

(c)



(d)

(e) (f)

Figure 2: Qualitative prediction results of the proposed model showing classification of different oral conditions: (a) Non-cancer (normal), (b) cancer, (c) lip cancer, (d) tongue cancer, (e) buccal cancer, and (f) OPMD (pre-cancerous condition).

The qualitative results in Fig. 2 demonstrate the model's ability to classify multiple oral conditions accurately. Image (a) shows a normal case correctly identified based on smooth texture and uniform colour. Image (b) represents a general cancer case with visible abnormalities, while (c) shows lip cancer with inflamed lesions. Image (d) identifies tongue cancer through texture variations, and (e) depicts buccal cancer with subtle pigmentation changes. Image (f) represents OPMD, where early-stage abnormalities are successfully detected. Overall, the results highlight the model's ability to distinguish between normal, cancerous, and pre-cancerous conditions effectively.

5.2 Loss Curve

The loss curve in Fig. 3 shows that both training and validation loss decrease steadily over epochs, indicating effective learning.

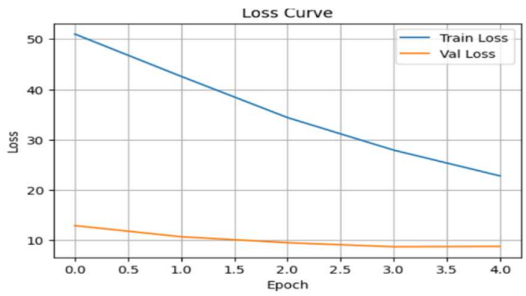


Figure 3 Loss Function

The validation loss follows a similar trend and remains lower than the training loss, suggesting good generalization without overfitting. Toward the final epochs, the validation loss stabilizes, indicating that the model is approaching convergence and further improvement is limited.

5.3 Accuracy Curve

The accuracy curve in Fig. 4 shows a steady improvement in both training and validation accuracy over epochs, indicating effective learning. The training accuracy (blue line) increases consistently from a lower starting point, while the validation accuracy (orange line) also improves and remains slightly higher in the early epochs. As training progresses, both curves come closer, showing that the model is learning generalized features without overfitting. Toward the final epochs, the validation accuracy stabilizes around 80%, suggesting that the model has reached near-optimal performance and further training may provide minimal gains.

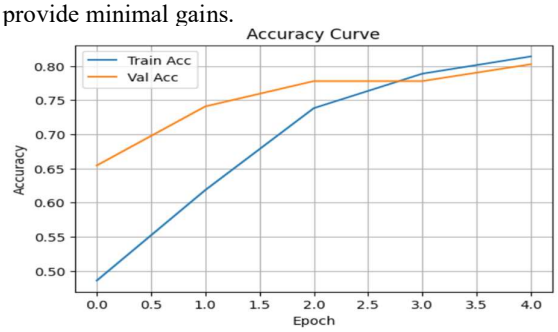


Figure 4 Accuracy Curve

5.4 Confusion Matrix

The confusion matrix in Fig. 5 illustrates the classification performance of the model across four classes: Buccal Cancer, Lip Cancer, OPMD, and Tongue Cancer. The diagonal values represent correct predictions, where the model performs best for OPMD with all 20 samples correctly classified,

indicating perfect detection. Lip Cancer and Tongue Cancer also show strong performance with 16 correct predictions each, although a few samples are misclassified as Buccal Cancer. Buccal Cancer shows comparatively lower performance, with 13 correct predictions and misclassifications into Lip Cancer and Tongue Cancer, suggesting some confusion due to similar visual features. Overall, the matrix indicates that the model performs well across most classes, with minor confusion primarily involving Buccal Cancer, while achieving excellent accuracy for OPMD.

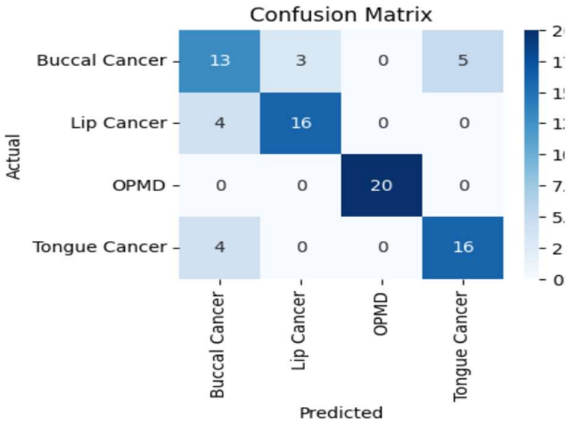


Figure 5 Confusion Matrix

5.5 Class Wise Performance

The class-wise performance chart in Fig. 6 shows precision, recall, and F1-score for each class. The model performs best on OPMD with perfect scores (1.0), indicating highly accurate detection. Lip Cancer and Tongue Cancer show good performance with scores around 0.78–0.85. Buccal Cancer has the lowest scores (around 0.62–0.63), suggesting some difficulty in classification. Overall, the model performs well across most classes, with room for improvement in Buccal Cancer.

Figure 6 Class Wise Performance

5.6 Grad-Cam Visualization

The Fig 7 shows the original image and its Grad-CAM visualization. The lesion on the lip is clearly visible in the original image, and the Grad-CAM heatmap highlights the same region as the most important for prediction. This indicates that the model focuses on relevant abnormal areas, demonstrating accurate localization and good interpretability.

Figure 7 Grad-CAM Visualization Highlighting Cancer-Affected Region in Oral Image

5.7 Impact of Multimodal Learning

The inclusion of patient metadata significantly improved the overall performance of the model. Compared to a CNN-only approach, the multimodal framework achieved higher classification accuracy, demonstrated better discrimination of early-stage cases such as OPMD, and showed enhanced robustness across diverse samples. These results indicate that integrating clinical information alongside image features strengthens the model's diagnostic capability and aligns closely with real-world medical practices, where both visual examination and patient history are essential for accurate diagnosis.

5.8 Effectiveness of Hybrid CNN–RNN Architecture

The integration of an RNN with the CNN enhanced the model's ability to capture contextual relationships within extracted features. This led to improved feature representation, better handling of complex lesion patterns, and higher classification performance compared to standalone CNN models. The hybrid architecture was particularly effective in identifying variations in early-stage lesions, contributing to more accurate and reliable predictions.

5.9 Region-Aware Segmentation Results

The segmentation module effectively identified regions of interest (ROI) corresponding to lesion areas, with generated masks closely aligning with the visual boundaries of affected regions. This contributed to improved classification accuracy, enhanced clinical interpretability, and better support for region-based diagnosis. These findings highlight the importance of incorporating segmentation into the diagnostic pipeline for more precise and clinically relevant analysis.

5.10 Explainability Analysis (XAI)

Grad-CAM visualizations provided meaningful insights into the model's decision-making process, with heatmaps consistently highlighting lesion regions. This confirms that the model focuses on clinically relevant areas, thereby improving model transparency, enhancing trustworthiness in clinical settings, and enabling validation by medical experts.

5.11 Comparative Discussion

Compared to existing approaches, the proposed model offers several key advantages, including the incorporation of multimodal data, the use of a hybrid CNN–RNN architecture, integration of region-aware segmentation, and the inclusion of Explainable AI techniques. These enhancements collectively contribute to improved accuracy, more robust feature representation, and better generalization across diverse samples.

5.12 Limitations

Despite its strong performance, the proposed study has certain limitations. It relies on publicly available datasets, and the use of synthetic metadata may not fully capture real clinical variability. Additionally, slight misclassification between OPMD and cancer was observed. These limitations highlight important areas for future research and improvement.

The experimental results demonstrate that the proposed multimodal hybrid framework significantly outperforms conventional models in terms of accuracy, robustness, and interpretability. The integration of segmentation and XAI further enhances clinical applicability, making the system suitable for real-world oral cancer diagnosis. The results confirm that the proposed multimodal hybrid CNN–RNN model provides superior performance and interpretability, making it a promising solution for early oral cancer detection.

6. CONCLUSION

This study presents an intelligent multimodal deep learning framework for the early diagnosis and detection of oral cancer. The proposed system integrates multi-source diversified datasets, multimodal feature fusion, a hybrid CNN–RNN architecture, region-aware segmentation, and Explainable Artificial Intelligence (XAI) into a unified pipeline. Unlike conventional approaches that rely on single-modal data and binary classification, the proposed model effectively combines visual features and patient metadata to perform multi-class classification (normal, OPMD, and cancer) along with localization of cancer-affected regions, thereby enhancing clinical relevance. The experimental results demonstrate that the proposed hybrid multimodal model achieves superior performance in terms of accuracy, robustness, and generalization compared to traditional CNN-based methods. The inclusion of metadata improves diagnostic capability, while the hybrid CNN–RNN architecture enables effective learning of both spatial and contextual features. Furthermore, the integration of segmentation allows region-aware analysis, which is essential for precise diagnosis and treatment planning. The incorporation of Explainable AI techniques, such as Grad-CAM, further enhances interpretability by providing visual explanations, thereby improving trust and usability in clinical settings. While the proposed framework demonstrates promising results, several avenues remain for future enhancement. The integration of real clinical datasets and electronic health records (EHR) can further improve reliability and practical applicability. Additionally, optimizing the model for real-time deployment on mobile or edge devices can enable accessible and low-cost screening, particularly in resource-limited environments. The use of advanced segmentation architectures and transformer-based models may further enhance localization accuracy. Moreover, incorporating self-supervised and federated learning approaches can address challenges related to limited labelled data and data privacy. Finally, extensive clinical validation with domain experts is essential to ensure real-world adoption and effectiveness. The proposed multimodal hybrid deep learning framework provides an accurate, interpretable, and clinically viable solution for early oral cancer diagnosis, with significant potential for future advancements in intelligent healthcare systems.

Reference

1. Devindi, G. A. I., et al. "Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer." *IEEe Access* 12 (2024): 124375-124390.
2. Gümele, Kaan, and Muhammet Sinan Başarslan. "Oral cancer classification with CNN based state-of-the-art transfer learning methods." *Black Sea Journal of Engineering and Science* 8.1 (2025): 94-101.
3. Amorim, Marcio Luis Munhoz, et al. "Early Cancer Detection Biosensors: Present Situation and Future Outlooks." *IEEE Sensors Reviews* (2025).
4. Ramya, Singaraju, R. I. Minu, and K. T. Magesh. "Xception spiking fractional neural network for oral squamous cell carcinoma classification based on histopathological images." *IEEE Access* (2025).
5. Patel, Saraswati, and Dheeraj Kumar. "Predictive identification of oral cancer using AI and machine learning." *Oral Oncology Reports* 13 (2025): 100697.
6. Patel, Saraswati, and Dheeraj Kumar. "Predictive identification of oral cancer using AI and machine learning." *Oral Oncology Reports* 13 (2025): 100697.
7. Thurnheer, Simon E., et al. "Early-stage oral cavity cancer (T1N0) with lymphatic drainage to the retropharyngeal lymph node: A therapeutic challenge." *Oral Oncology Reports* 13 (2025): 100713.
8. Bogaert, Brenda. "Oral potentially malignant disorders: Challenges for patient participation due to opacity." *Oral Oncology Reports* 13 (2025): 100731.
9. Shashikala, K. S., and S. Umamaheswaran. "Artificial Intelligence in Oral Health: A Study

for the Detection of Oral Precancerous Conditions." *2025 IEEE International Conference on Advances in Computing Research On Science Engineering and Technology (ACROSET)*. IEEE, 2025.
10. Kumar, G. Manoj, et al. "Early Stage Oral Cancer Detection using Machine Learning and Deep Learning Algorithms." *2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 2025.
11. Sivakumar, R., and K. Sathish. "Image-Based Oral Cancer Detection with Advanced Deep Learning Methodologies." *2025 Control Instrumentation System Conference (CISCON)*. IEEE, 2025.
12. Vennila, G., et al. "Temporal Analysis of Oral Lesions Using RNN for Early Detection of Oral Cancer." *2025 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*. IEEE, 2025.
13. Ashwini, A., et al. "Bio inspired optimization techniques for disease detection in deep learning systems." *Scientific Reports* 15.1 (2025): 18202.
14. Ayyapa, Valluri, et al. "Non-invasive oral cancer detection using hyperspectral imaging and advanced spectral unmixing models." *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*. IEEE, 2024.
15. Jagadesh, T., et al. "Oral cancer detection using convolutional neural networks." *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*. IEEE, 2024.
16. Ghosh, Joita, Khushi Mishra, and M. Sindhuja. "Oral Vision: Advanced Image Recognition for Oral Cancer." *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*. IEEE, 2024.
17. Cherukupalli, Maitreya, et al. "Deep Learning-Based Oral Cancer Detection through Speech Analysis." *2024 Control Instrumentation System Conference (CISCON)*. IEEE, 2024.
18. Kumar, BH Manjunatha, et al. "Advanced meta-heuristic algorithm based on AlBiruni earth radius and particle swarm optimization methods for oral cancer detection." *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*. Vol. 1. IEEE, 2024.
19. Sahoo, Rakesh Kumar, et al. "Diagnostic performance of artificial intelligence in detecting oral potentially malignant disorders and oral cancer using medical diagnostic imaging: a systematic review and meta-analysis." *Frontiers in oral health* 5 (2024): 1494867.
20. Devindi, G. A. I., et al. "Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer." *IEEe Access* 12 (2024): 124375-124390.
21. Das, Madhusmita, et al. "An Ensemble deep learning model for oral squamous cell carcinoma detection using histopathological image analysis." *IEEE Access* 12 (2024): 127185-127197.
22. Goswami, Bibek, et al. "Classification of oral cancer into pre-cancerous stages from white light images using LightGBM algorithm." *Ieee Access* 12 (2024): 31626-31639.
23. Goswami, Bibek, et al. "Classification of oral cancer into pre-cancerous stages from white light images using LightGBM algorithm." *Ieee Access* 12 (2024): 31626-31639.
24. Jangde, Keshika, and Ranu Pandey. "Deep Learning for Oral Cancer Detection Using ResNet-50, Bi-LSTM, and Multimodal Fusion." *Vascular and Endovascular Review* 7.2 (2024): 88-99.
25. Verma, Damini, et al. "An Ultrasensitive Electrochemical Immunosensor Comprising Green Synthesized α-Fe 2 O 3 NPs_rGO Nanocomposite for Determination of Oral Cancer." *IEEE Sensors Letters* 7.12 (2023): 1-4.
26. Nourinovin, Shohreh, et al. "Terahertz characterization of ordinary and aggressive types of oral squamous cell carcinoma as a function of cancer stage and treatment efficiency." *IEEE Transactions on Instrumentation and Measurement* 72 (2023): 1-9.
27. Dixit, Shriniket, Anant Kumar, and Kathiravan Srinivasan. "A current review of machine learning and deep learning models in oral cancer diagnosis: recent technologies, open challenges, and future research directions." *Diagnostics* 13.7 (2023): 1353.
28. Chakraborty, Parnasree, et al. "Artificial Intelligence-based Oral Cancer Screening System using Smartphones." *Engineering, Technology & Applied Science Research* 13.6 (2023): 12054-12057.
29. Haq, Ihtisham Ul, et al. "Unveiling the future of oral squamous cell carcinoma diagnosis: an innovative hybrid AI approach for accurate histopathological image analysis." *IEEe Access* 11 (2023): 118281-118290.
30. Myriam, Hadjouni, et al. "Advanced meta-heuristic algorithm based on Particle Swarm and Al-biruni Earth Radius optimization methods for oral cancer detection." *IEEE Access* 11 (2023): 23681-23700.
31. Hassan, Mohamed Abul, et al. "Anatomy-specific classification model using label-free

FLIm to aid intraoperative surgical guidance of head and neck cancer." *IEEE Transactions on Biomedical Engineering* 70.10 (2023): 2863-2873.

32. Pathuthara, Aqil, et al. "Transfer learning-based model for oral cancer detection." *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*. IEEE, 2023.
33. Myriam, Hadjouni, et al. "Advanced meta-heuristic algorithm based on Particle Swarm and Al-biruni Earth Radius optimization methods for oral cancer detection." *IEEE Access* 11 (2023): 23681-23700.
34. Wang, Ning, Yuan Liu, and Hongwen Li. "An Efficient and Fast, Noninvasive, Auto-Fluorescence Detection Method for Early-Stage Oral Cancer." *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 1-11.
35. Shamim, Mohammed Zubair M. "Hardware deployable edge-AI solution for prescreening of oral tongue lesions using TinyML on embedded devices." *IEEE Embedded Systems Letters* 14.4 (2022): 183-186.
36. Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71.3 (2021): 209-249.
37. Lin, Huiping, et al. "Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis." *Journal of Biomedical Optics* 26.8 (2021): 086007-086007.
38. Sharma, Neha, et al. "Multifractal texture analysis of salivary fern pattern for oral pre-cancers and cancer assessment." *IEEE Sensors Journal* 21.7 (2021): 9333-9340.
39. Welikala, Roshan Alex, et al. "Automated detection and classification of oral lesions using deep learning for early detection of oral cancer." *Ieee Access* 8 (2020): 132677-132693.