Early Detection of Oral Cancer

Leveraging Multimodal Machine Learning to Mirror the Real-World Diagnostic Workflow

Fadwa F. Elfeituri
Biomedical Graduate Education, Georgetown University
HIDS-7006: AI for Health Applications

May 2025

1. Background and Significance

Oral cancer is a malignant neoplasm arising in the lip or oral cavity, most commonly manifesting as Oral Squamous Cell Carcinoma (OSCC) which accounts for over 90 percent of cases. It encompasses malignancies of the mucosal lip, tongue, gum, floor of the mouth, and hard and soft palate.[1]

Globally, oral cancer remains the sixth most common cancer, with 389,846 new cases and 188,438 deaths reported in 2022.[2] In the United States, the American Cancer Society estimates 59,660 new cases and 12,770 deaths from oral cavity and pharyngeal cancers in 2025, representing 2.9 percent of all new cancer diagnoses.[3]

Early detection is impeded by the fact that oral cancers often present with subtle symptoms, such as small white or red patches, mild ulcerations, or painless lumps, that can mimic common benign or reactive conditions. Such heterogeneity frequently leads to misclassification or delayed clinical suspicion, placing substantial reliance on specialist expertise and thorough inspection.[4]

Despite therapeutic advances, the majority of patients – approximately 70 percent – are still diagnosed at advanced stages, at which point the five-year survival rate plummets from 83.7 percent for localized disease to 38.5 percent for metastasized cancer. [4] The current diagnostic gold standard – scalpel biopsy by histopathologic evaluation – is invasive, time-consuming, are prone to significant inter- and intra-observer variability, often delaying definitive diagnosis and treatment.[5] In low-resource and rural settings, these delays are exacerbated by limited specialist access and diagnostic infrastructure, contributing to poorer outcomes.[6]

To address these challenges, researchers have explored machine learning (ML and deep learning (DL)-based approaches for non-invasive lesion classification. A recent systematic review and meta-analysis reported a summary SROC of 0.94 for AI-assisted detection of oral precancerous and cancerous lesions, with sensitivity 89.9 percent and specificity 89.2 percent.[7] Individual studies using convolutional neural networks, support vector machines, and ensemble classifiers have demonstrated classification accuracies ranging from 80.9 percent to 98 percent.[8] Moreover, point-of-care implementations – such as smartphone-based screening apps – have achieved high diagnostic performance in pilot settings, highlighting the feasibility of rapid, bedside assessments. However, despite these results, most existing approaches rely on a single data modality – typically clinical photographs or histopathology images – and are validated on limited datasets, limiting generalizability.[9]

2. Project Objective

The overarching goal was to develop a multimodal AI-driven diagnostic tool that fuses four complementary data streams—oral clinical photographs, histopathology slides, radiographic images, and salivary biomarker profiles—into a unified, clinician-friendly system for early oral cancer detection. However, due to time constraints, the current work concentrated on creating a custom convolutional neural network (CNN) to perform binary classification of clinical photographs (cancer vs. non-cancer). This proof-of-concept model was trained, validated, and tested on a harmonized image dataset, laying the groundwork for subsequent integration of additional modalities.

3. Methods

3.1. Data Source

The study utilized a binary-labeled corpus of clinical oral photographs sourced from two publicly available Kaggle repositories:

- Oral Cancer Dataset [10]: Comprised 950 images (500 malignant, 450 benign) acquired via intraoral digital photography in tertiary care centers. Lesion labels were assigned based on histopathological confirmation.
- Oral Cancer (Lips and Tongue) Images [11]: Included 131 images (87 malignant, 44 benign) collected in ENT outpatient clinics and annotated by board-certified oral pathologists.

The dataset was constructed by merging the two repositories where each raw image served as one distinct example, with class membership determined by its parent folder ("Cancer" or "Non-Cancer")

3.2. Data Preprocessing

A comprehensive preprocessing pipeline was implemented to ingest raw images, enforce quality control, and build a richly annotated manifest. The process included:

1. **File Identification:** Root folders were scanned using the Python Imaging Library (PIL) to identify the raw root directories and collect every image file path into a master list.
2. **Integrity Check:** Each file in the raw directories was opened with PIL and excluded any corrupted or unreadable images.
3. **Image Conversion and Resizing:** Valid images were converted to RGB to ensure consistency and resized to 128 x 128 pixels. Each processed image was saved as an optimized JPEG (quality level 80) under a new directory structure organized by class.
4. **Metadata Extraction and Manifest Creation:**
    - For each image, key attributes were extracted including original and resized dimensions (width and height), color mode, file format, and file size in kilobytes. Class labels were encoded as (0 = Non-Cancer, 1 = Cancer).
    - These details were compiled into a manifest CSV file that linked each image to its metadata and binary label.
5. **Feature Engineering:** The manifest file was further enriched with engineered features to improve model interpretability and performance:
    - **Aspect Ratios:** Calculated for both original and resized images as width divided by height.
    - **Scale Factor:** Computed as the ratio of the original image area to the resized image area.
    - **One-Hot Encoding:** Categorical variables such as color mode and file format were converted into binary indicator columns.
    - **Boolean Conversion:** All Boolean-type columns were cast to numeric values (0 or 1)
    - **Normalization:** Continuous variables including file size, image dimensions, aspect ratios, and scale factor, were standardized using z-score normalization.

    The final output included a structured feature matrix (X) and a binary outcome label (Y) that were used to train and evaluate the classification model.

3.3. Model Development

To classify oral cancer from clinical photographs, two convolutional neural network (CNN) models were developed using TensorFlow and Keras: (1) a multimodal two-branch model that integrates both image and metadata inputs, and (2) a follow-up image-only model. This design allowed for comparative analysis of how structured metadata contributes to classification performance.

3.3.1. Two-Branch Multimodal CNN

This approach focused on designing and optimizing a convolutional neural network (CNN) capable of learning from both image data and associated metadata, allowing it to capture visual and contextual patterns relevant to diagnosis.

1. **Data Pipeline and Preparation:** Image paths and structured metadata were loaded from the preprocessed manifest files. Unique file names were extracted to prevent data leakage during dataset splitting. A stratified split was applied based on the binary target (label_idx), allocating 70% for training and 30% as a temporary holdout. This holdout portion was further divided into validation (10%) and test (20%) subsets, preserving class distribution across all sets.
2. **Model Architecture:** A custom dual-input neural network architecture was implemented, combining a convolutional branch for image features with a dense branch for metadata.
    - **Image Branch:**
        i. Input: 128 × 128 × 3 RGB image tensor
        ii. Two Conv2D layers with 32 and 64 filters (3×3), each followed by BatchNormalization, ReLU activation, and MaxPooling
        iii. Flatten layer and Dropout (rate configurable via tuning)
    - **Metadata Branch:**
        i. Input: 7 structured features
        ii. Dense layer with 32 units and ReLU activation
        iii. Dropout layer (configurable)
    - **Fusion and Output:**
        i. Concatenation of image and metadata branches
        ii. Dense layer (units configurable via tuning) with ReLU activation
        iii. Output layer with 1 sigmoid-activated neuron for binary classification
3. **Model Optimization and Tuning:** The model was compiled with the Adam optimizer (learning rate = 0.0001) and trained with binary cross-entropy loss. Metrics used to evaluate performance included accuracy, precision, recall, and AUC. To improve generalization and prevent overfitting, the following strategies were applied:
    - **Regularization**: Dropout layers and L2 weight penalties were incorporated into both input branches. These parameters were optimized during tuning.
    - **Early Stopping & Model Checkpointing**: Training was monitored using validation loss with a patience of 3 epochs. The best model was saved based on validation AUC.
    - **Hyperparameters Tuning**: To identify the optimal model configuration, a structured hyperparameter tuning loop was implemented. Three tuning configurations were tested,

each varying the dropout rate, L2 regularization strength, and number of dense units:

```
# Specify Tuning Configurations
configs = [
    {'run_name': 'run_A', 'dropout': 0.3, 'l2': 0.0005, 'dense': 128},
    {'run_name': 'run_B', 'dropout': 0.4, 'l2': 0.0001, 'dense': 128},
    {'run_name': 'run_C', 'dropout': 0.5, 'l2': 0.0005, 'dense': 128},
]
```

Each configuration was trained for up to 15 epochs with early stopping and logged for comparison. The best-performing model (Run B) was selected based on validation accuracy and generalization to the test set.

3.3.2. Image-Only CNN Model

After training and tuning the two-branch model, an image-only CNN was developed to serve as a comparison baseline. This model excluded metadata entirely and relied only on visual features extracted from clinical photographs.

- Input: 128 × 128 × 3 image
- Two convolutional blocks (32 and 64 filters)
- Flatten → Dense(128, ReLU) → Dropout
- Output: Dense(1, sigmoid)

The same training configuration and evaluation criteria were used to compare this model against the two-branch model version.

4. Results

To evaluate model performance, a comprehensive pipeline was implemented that included:

- Classification reports summarizing precision, recall, F1-score, and class support
- Confusion matrices visualizing true vs. predicted labels
- Training and validation curves for accuracy and loss to assess learning stability and potential overfitting.

All evaluations were conducted on the held-out test set.

4.1. Performance Metrics

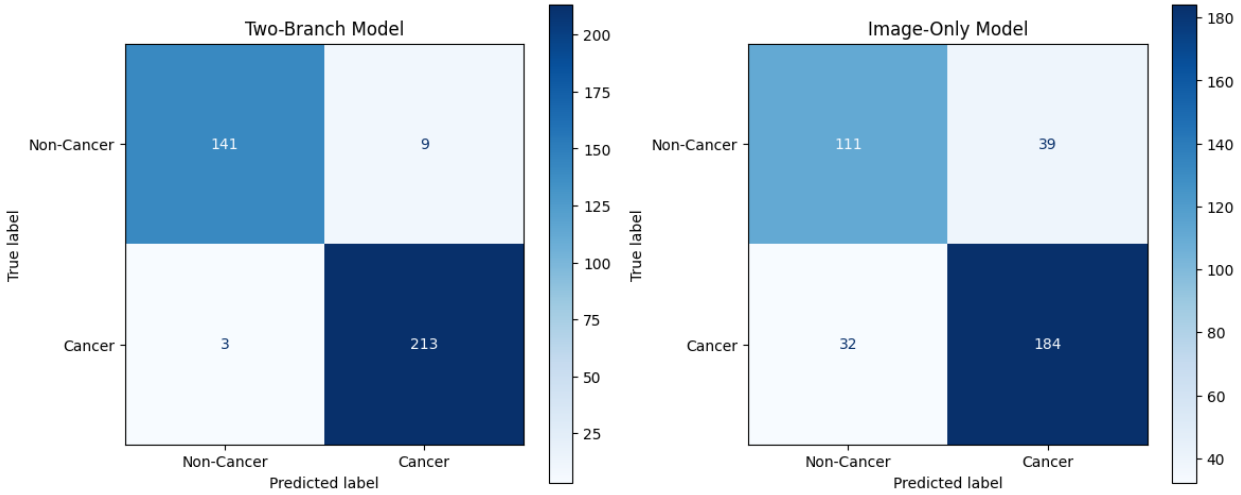
	Two-Branch Model		Image-Only Model	
Metric	Cancer	Non-Cancer	Cancer	Non-Cancer
Precision	0.96	0.98	0.83	0.78
Recall	0.99	0.94	0.85	0.74
F1-Score	0.97	0.96	0.84	0.76

The two-branch model significantly outperformed the image-only model in all metrics—particularly in recall, where it correctly identified 99% of cancer cases. The image-only model, while still functional, missed more positive cases and showed more false positives.

4.2. Confusion Matrix

The confusion matrix revealed that the two-branch model reduced both false negatives (missed cancer cases) and false positives (misclassified non-cancer), improving its clinical reliability.

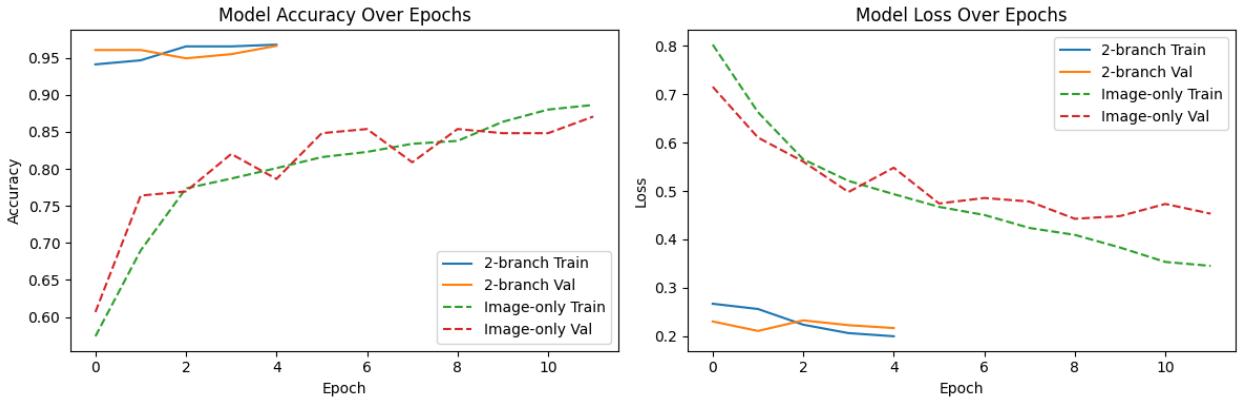
Figure 1. Confusion Matrix Comparison: Two-Branch vs. Image-Only CNN Models



4.3. Training and Validation Curves

The two-branch model achieved higher and more stable validation accuracy with consistently low loss, indicating better generalization. The image-only model showed fluctuating validation performance and a noticeable gap between training and validation loss, suggesting mild overfitting and reduced robustness.

Figure 2. Training and Validation Accuracy and Loss Curves



4.4. Prediction Accuracy and Misclassification Patterns

To qualitatively compare how both models performed on challenging cases, sample predictions were visualized alongside true labels. The two-branch model correctly identified all 10 cancerous lesions shown, while the image-only model misclassified 3 of them as non-cancer. These errors tended to occur in cases where lesions presented as less visually prominent or were located in areas prone to benign variation (e.g., inner lip or lateral tongue). This suggests that the additional metadata provided subtle cues that supported more confident classification, particularly in borderline cases. This qualitative

review aligns with the numerical findings: the multimodal model achieved higher recall and precision, reducing both false negatives and false positives.

Figure 3. Side-by-Side Comparison of Model Predictions vs. True Labels for the Two Models

4.5. Final Performance Summary of the Two Models

The two-branch model achieved a best validation accuracy of 96.6% and test accuracy of 96.7%, outperforming the image-only model, which reached 87.1% validation and 80.6% test accuracy. These results confirm the performance advantage of incorporating metadata into the classification pipeline.

5. Conclusion and Future Directions

This project highlights the potential of AI-driven diagnostic tools in addressing delays and subjectivity in oral cancer detection. By combining clinical images with structured metadata, the developed two-branch model significantly improved diagnostic accuracy. Such a model could be deployed as a clinical decision support tool, aiding non-specialists in the early identification of suspicious lesions, particularly in under-resourced settings. For example, it could be embedded in handheld digital oral scanners or mobile platforms, making diagnostic support accessible at the point of care.

Interpretability was partially addressed through metadata features like image format and dimensions, which likely provided indirect indicators of lesion prominence or image framing. However, the internal decision-making process remains largely opaque, and understanding exactly how the model arrives at its predictions continues to be a limitation.

Moving forward, the goal is to expand the system into a comprehensive multimodal diagnostic platform that mirrors real-world clinical workflows. This includes training on multiclass annotated datasets, integrating additional diagnostic inputs such as histopathological slides, radiographic images, and salivary biomarkers, and incorporating demographic and clinical information to enhance contextual accuracy. This vision aligns with the broader shift toward AI-augmented care—empowering clinicians with earlier, more accurate, and accessible diagnostic tools that can improve outcomes and reduce global disparities in oral cancer survival.

References

1. World Health Organization. (2025). Oral health. https://www.who.int/news-room/fact-sheets/detail/oral-health
2. Al-Ghamdi, S., Alkhaldi, H., & Arafah, M. (2024). Oral cancer: An overview of global burden, risk factors, and prevention strategies. Journal of Oral Biology and Craniofacial Research, 14(2), 170–176. https://doi.org/10.1016/j.jobcr.2024.01.013
3. American Cancer Society. (2025). Cancer facts & figures 2025. American Cancer Society. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2025/2025-cancer-facts-and-figures-acs.pdf
4. González-Moles, M. Á., Aguilar-Ruiz, M., & Ramos-García, P. (2022). Challenges in the early diagnosis of oral cancer, evidence gaps and strategies for improvement: A scoping review of systematic reviews. *Cancers, 14*(19), 4967. https://doi.org/10.3390/cancers14194967
5. Warnakulasuriya, S. (2018). Clinical features and presentation of oral potentially malignant disorders. Oral and Maxillofacial Surgery Clinics of North America, 25(4), 467–484. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5980879/
6. Zahnd, W. E., Fogleman, A. J., Jenkins, W. D., Watanabe-Galloway, S., Boucher, K. M., Proctor, E. K., & Vanderpool, R. C. (2022). Rural–urban disparities in cancer outcomes: Opportunities for future research. *CA: A Cancer Journal for Clinicians, 72*(2), 167–190. https://doi.org/10.3322/caac.21721
7. Brizuela, S., Chieng, H. C., & Milinovich, A. (2024). Diagnostic accuracy of artificial intelligence in detecting oral cancer: A systematic review and meta-analysis. *International Journal of Surgery, 116*, 419–429. https://doi.org/10.1097/IJ9.0000000000000521
8. Shaik, S., Gopinath, P., & Rao, A. (2024). Artificial intelligence in oral cancer: A comprehensive scoping review of diagnostic and prognostic applications. *Journal of Oral Pathology & Medicine*. Advance online publication. https://www.researchgate.net/publication/388369246
9. Uthoff, R. D., Thomas, S. S., Bagwell, J., Arnot, A. R., Ceballos, D. R., Kuriakose, M. A., ... & D'Souza, G. (2018). Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device using both autofluorescence and reflectance imaging. *PLOS ONE, 13*(12), e0207493. https://doi.org/10.1371/journal.pone.0207493
10. Zaidpy. (n.d.). *Oral cancer dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/zaidpy/oral-cancer-dataset/data
11. Shivam17299. (n.d.). *Oral cancer lips and tongue images* [Data set]. Kaggle. https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images