10.21608/adjc.2025.378150.1754 Advanced Dental Journal
Volume (7) | Issue 4 | Pages 627- 641 ©Faculty of Dentistry-Cairo University

Original Article

The Application of Artificial-based Models to Classify Oral Cavity Findings Based on Clinical Image Analysis

Noran Ayman M. Ismael [1], Fat'heya M. Zahran [1], Yomna Safaa EL-Din [2], Noha Adel Azab [1]

[1] Department of Oral Medicine and Periodontology, Faculty of Dentistry, Cairo University, Cairo, Egypt.
[2] Department of Oral Medicine and Periodontology, Faculty of Dentistry, Cairo University, Cairo, Egypt.

E-mail: noran.ayman@dentistry.cu.edu.eg
Submitted: 23-4-2025
Accepted: 3-6-2025

Abstract

Objective: This study aims to collect a comprehensive and versatile dataset to provide a solid foundation to assess the performance of five promising models for early detection of oral cancer.

Materials and methods: Clinical photographs of the entire mouth were obtained from patients visiting the Oral Medicine clinic at Cairo University's Faculty of Dentistry. These images were labeled and prepared to evaluate five CNN models, examining various data processing methods. The study incorporated augmentation techniques for all models and tested each model both with and without oversampling.

Results: The dataset comprises 5, 616 intraoral images, which are subdivided according to the presence and type of oral lesion. These include 2,686 images classified as 'Normal,' 1,410 as 'Benign and inflammatory,' and 1,520 as 'Potentially malignant and malignant.' The findings indicate that oversampling (V1) significantly improved model performance, particularly for GoogleNet, which consistently ranked among the top models in precision, accuracy, recall, and F1-score. InceptionResNetvr2 performed better in all evaluation metrics without oversampling. EfficientNet b4 showed similar results with and without oversampling, while ViT was the least consistent.

Conclusion: These results highlight the dataset's variability and complexity, revealing challenges in processing large-scale clinical oral images. This advances versatile models for diverse diseases.

Keywords: Malignant, Oral Cancer, potentially malignant, Deep learning, Intraoral images

I. INTRODUCTION

Head and neck squamous cell carcinomas (HNSCC) are a leading cause of cancer-related mortality, with oral cancer accounting for 5.8% of new cases in males and 2.3% in females according to the Globocan (Bray et al., 2024; Ferlay et al., 2015). Cases projected to rise by 62% to 856,000 by 2035 (***Shield et al., 2017)***. Late-stage diagnosis is often due to limited awareness and healthcare access and results in over two-thirds of cases being identified after metastasis, contributing to high mortality rates ***(Kavyashree, Vimala and Shreyas, 2024; Ilhan et al., 2020)***. Advances in medical image analysis using artificial intelligence (AI) are innovative tools for diagnosis and treatment planning which offer promising solutions to these challenges ***(Warin et al., 2022)***.

Deep learning (DL), a subset of AI, builds algorithms by layering simple notions on top

of each other ***(Ferreira, Silva and Valente, 2021)***. A major advantage of DL is its ability to automatically extract features from raw data, eliminating the need for manual feature engineering ***(Klang, 2018)***. Convolutional neural networks (CNNs), a type of supervised DL model, are highly effective for image analysis and classification, as they learn directly from raw pixels to final classifications without requiring manual feature extraction ***(Ghaffar Nia, Kaplanoglu and Nasab, 2023; Han, Liu and Fan, 2018)***.

Deep learning approaches generally rely on large datasets to train models effectively and avoid overfitting, a situation where model performs well on the training data but fails to generalize to new, unseen data ***(Chlap et al., 2021)***. DL models trained on extensive datasets allow their learned weights to be transferred to other models for testing or further training, a process known as 'transfer learning' ***(Krishna and Kalluri, 2019)***.

Transfer learning improves task performance by utilizing knowledge gained from solving related tasks in different contexts but within a shared domain. This approach enables the model to build on existing knowledge, rather than starting from scratch ***(Hosna et al., 2022).***

The absence of publicly accessible datasets has limited researchers' ability to develop AI algorithms that could assist general practitioners in the early detection of oral cancer ***(Sengupta et al., 2022)***. Previous studies on the AI-based early detection of oral cancer have relied on limited datasets, with a maximum of 2160 unedited images, and have consistently recommended the use of larger datasets to improve model performance ***(Lee et al., 2022; Rashid et al., 2024; Tiryaki et al., 2024; Welikala et al., 2020)***.

Additionally, small datasets usually suffer from class imbalance, which hinder AI-based data analysis. This imbalance arises because one class often significantly outnumbers others due to insufficient patient availability for certain diseases, high costs, privacy and security concerns, and data complexity ***(Abd Elrahman and Abraham, 2013; Goceri, 2023)***. AI models trained on such imbalanced data tend to produce biased classifiers, disproportionately favoring the majority class and compromising the detection of minority classes, which are often clinically critical ***(Rajaraman, Ganesan and Antani, 2022; Larrazabal et al., 2020)***. These obstacles are usually managed by employing data augmentation and oversampling techniques ***(Goodfellow, Bengio and Courville, 2016)***.

Oversampling involves increasing the number of minority class instances by randomly replicating them, thereby balancing the class distribution ***(Fernández et al., 2018)***. Oversampling enhances model performance by increasing training samples and improving generalization, reducing overfitting risks. However, it can also cause overfitting if the generated data lacks diversity, leading the model to memorize existing patterns instead of learning new ones ***(Jalata, Khan and Nakarmi, 2024)***.

Our aim was to collect a comprehensive and versatile dataset to provide a solid foundation to assess the performance of five promising models for early detection of oral cancer. Our secondary outcome was to assess various data processing techniques that can enhance the predictive performance of these chosen models.

II. MATERIALS AND METHODS

This study was conducted at the Oral Medicine Clinic, Faculty of Dentistry, Cairo University, between March and December 2024. Ethical clearance was granted by the Faculty's Research Ethics Committee (Approval No. 6-3-24). A block diagram of the methodology of the study is shown in figure 1.

Dataset collection and labeling

The dataset consists of 5,616 intraoral images in JPG format, with varying dimensions depending on the capture device, whether a smartphone or a dedicated camera. The images were classified by an oral medicine specialist according to the presence and type of oral lesions. The classification included 2,686 images in the 'Normal' category, 1410 images

in the 'Benign and inflammatory' category, and 1520 images in the 'Potentially malignant and malignant' category. Additionally, each category was organized into 11 distinct anatomical sites. With 3 subcategories per anatomical site, this yielded 33 possible combinations. A representative sample is shown in figure 2.

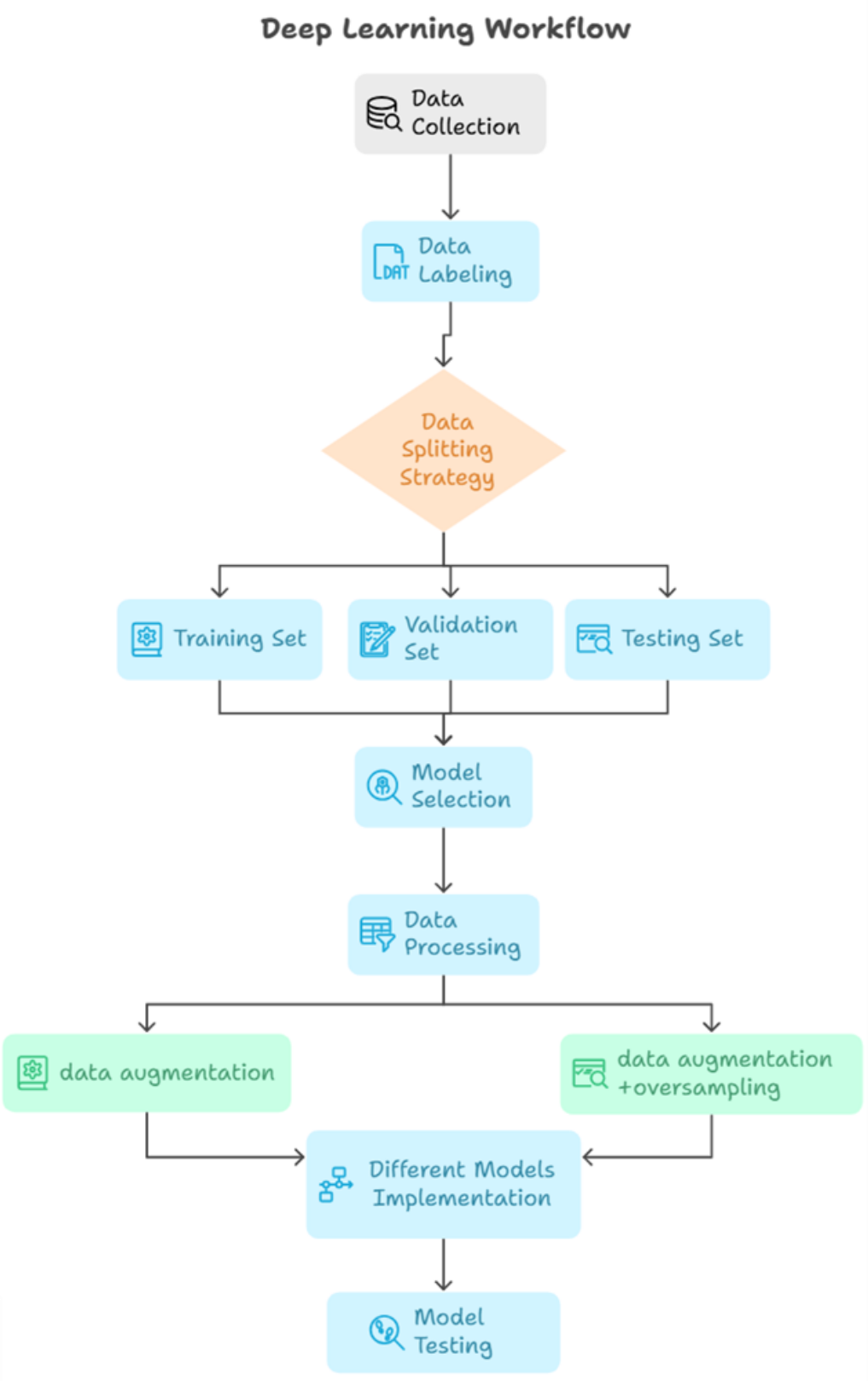

Figure 1: The deep learning workflow

Figure 2: A representative sample of the three categories: Normal; Benign and inflammatory; and Potentially malignant and malignant

Data Splitting Strategy:

Stratified sampling was employed to guarantee that each subset (Training, validation and testing) contains approximately the same percentage of each category as in the original dataset to avoid class imbalances. The dataset was split into three distinct subsets: a learning set for training the model, a validation set used for tuning the hyperparameters and monitoring the performance, and a testing set reserved exclusively for the final evaluation of the model, as shown in Table 1.

Model Selection and Transfer Learning:

We initially evaluated 17 pretrained models on ImageNet to identify the most effective architecture for our task. Based on their performance and computational efficiency, we selected the five promising models for further experimentation shown in Table 2. The number of parameters in deep learning refers to the internal variables learned from training data, which define how the model processes inputs and makes predictions. These parameters are adjusted during training to minimize errors. More parameters increase model complexity, requiring greater computational power and larger datasets, but can also lead to overfitting ***(Alpaydin, 2020).***

We explored multiple methods to integrate site information and image data for cancer

Table 1: Distribution of data among different phases of model development and evaluation according to lesion category

	Learning (70%)	Validation (15%)	Testing (15%)	Total
Normal	1880	403	403	2,686
Benign and inflammatory	986	212	212	1410
Potentially malignant and malignant	1,064	228	228	1520

Table 2: The selected five models for our task and their description

Model	Description	Number of parameters
Vision Transformer (ViT):	A transformer-based model that processes image patches as sequences, leveraging self-attention for capturing long-range dependencies ***(Alexey, 2020)***.	86 million
EfficientNet-B4 (NVIDIA):	A CNN that achieves optimal performance by proportionally scaling depth, width, and resolution ensuring high computational efficiency across various hardware platforms ***(Tan and Le, 2019)***.	19 million
Resnet 50:	A CNN based on 'residual blocks', which improves gradient flow, enabling deeper network training by minimizing the vanishing gradient problem—where gradients become too small to update earlier layers effectively—ensuring stable and efficient learning ***(He et al., 2016).***	25.5 million
GoogleNet (Inception V1):	A CNN based on 'inception modules', which allow the network to capture features at multiple scales using parallel convolutions within the same layer ***(Szegedy et al., 2014).***	19.5 million
Inception-ResNet-V2	A hybrid architecture that integrates Inception modules with residual connections. This allows the architecture to benefit from both models, enabling faster training along with skipping some layers during training, which makes it efficient for feature extraction in complex medical images ***(Neshat et al., 2024).***	55.8 million

classification in the previously mentioned models as shown in Figure 3:

A. Image-Based Classifier:

In this approach, only image features are used to predict the cancer class (Potentially malignant and malignant, Benign and inflammatory, normal). Site information is incorporated during the decision process using masking-based classification.

Masking-Based Classification: One architectural extension involved leveraging site information to improve classification accuracy. Each model was trained to predict all combinations of risk levels and site locations (3

× number of sites). During inference, a mask was applied to the final logits to retain only the three probabilities corresponding to the known site of the input image. This masking mechanism ensured that predictions were site-specific, which improved performance by reducing unnecessary cross-site confusion.

B. Site-Enhanced Classifier:

A separate encoder processes the site information (one-hot encoded or embedded) and concatenates these features with the image features before the classification layer. This allows the model to use both visual and location-based data.

C. Early Site Integration:

In this architecture, site information is introduced earlier in the process by encoding the site and adding the resulting features to the image before being passed through the image encoder. This allows for site-specific patterns to be learned throughout the entire model.

Figure 3: Different methods to integrate site information presented in 1. Image based classifier with masking, 2. Site-enhanced classifier, and 3. Early site integration

In order to assess the impact of the oversampling technique, each selected model was trained and tested using three approaches: Version 1 (V1), utilized on-the-fly data augmentation and oversampling; Version 2 (V2) utilized on-the-fly data augmentation without oversampling; while Version 3 (V3) utilized on-the-fly data augmentation, without oversampling, with the addition of combining contralateral mirror sides, e.g.: right and left buccal mucosa.

Data processing:

- **On-the-fly data augmentation techniques:**

On-the-fly data augmentation is applied to enhance the variability and robustness of the model; it's done dynamically during the training process rather than preprocessing and storing augmented data in advance. The augmentation parameters were empirically chosen to ensure the augmented images were still clinically meaningful. These techniques included:

1. Random Resized Cropping: Each image was randomly cropped and resized to a fixed shape, with a scaling factor sampled from a range of 0.8 to 1.0. This simulates zoom-in and zoom-out effects.

2. Random Rotation: Images were randomly rotated within a range of ±35◦. This augmentation helps the model generalize to rotated variations of the input.

3. Random Affine Transformations: Affine transformations were applied, allowing for small random rotations (±10◦), translations (up to 15% in both horizontal and vertical directions), and shearing up to 20%. These transformations introduce geometric variability, improving spatial robustness.

4. Color Jittering: The brightness of each image was randomly adjusted within a range of 50% to 100% of its original value. This addresses variations in lighting conditions in the dataset. Each augmentation technique was carefully chosen to introduce realistic variations in the data while preserving essential features necessary for accurate model performance.

- Oversampling technique:

The aim of trying this technique in V1 is to assess its impact on the results. It was done as follows:

Identifying the Maximum Class Size:

We first identified the site-risk level combination with the largest number of images in the dataset—the left buccal mucosa in the Potentially malignant and malignant category, containing 725 images. We used this maximum as the target sample size for balancing all 33 combinations (derived from 3 risk categories across 11 anatomical sites).

Data Augmentation and Oversampling:

To address class imbalance in combinations with fewer images, we performed oversampling by duplicating existing samples, ensuring equal representation across all site-risk level categories. This approach prevented model bias toward majority classes while maintaining balanced exposure to each risk level. To counteract potential overfitting from duplicated images, we applied randomized data augmentation techniques during each training epoch, introducing variability that forced the model to learn generalizable features rather than memorizing repeated data. Finally, during image preprocessing, all augmented samples were resized to meet each model's input specifications: 224×224 pixels for GoogleNet (InceptionV1), ViT, and ResNet50; 380×380 for EfficientNet B4; and 299×299 for InceptionResNetV2.

Model Tuning:

All experiments were conducted on Kaggle's cloud-based platform using their free GPU resources. Model tuning is the process of optimizing hyperparameters for effectiveness ***(Alpaydin, 2020)***

cloud-based platform using their free GPU resources. Model tuning is the process of

optimizing hyperparameters for effectiveness ***(Alpaydin, 2020).***

Hyperparameters are predefined settings that shape the training process and model structure, remaining fixed while parameters evolve during training. Key hyperparameters include:

- *Batch size*, which determines the number of samples processed simultaneously.
- *Dropout rate*, which enhances model's generalization by randomly omitting a percentage of the hidden units during training, avoiding co-adaptation of feature detectors which reduces overfitting. The dropout rate determines the fraction of neurons that are randomly dropped during training (Shen et al., 2017).
- *Learning rate*, which controls the speed of parameter updates. Smaller batches may converge faster but require a lower learning rate to avoid overshooting optimal minima (the point of minimal error), while larger batches can achieve better minima but lack the regularization effect of smaller batches due to their lower variance (Goodfellow, Bengio and Courville, 2016). Thus, it's recommended in medical image classification to choose small batch size (usually 32 to 64) with a low learning rate (Kandel and Castelli, 2020)

During the training process, the models were trained for a maximum of 40 epochs and the learning rate was 0.001. We experimented with batch sizes of 32, 58 and 128 but no significant differences in the model performance were observed across this range, consequently, an average batch size of 64 was selected. To optimize model performance, several techniques were fine-tuned, including early stopping, scheduled learning rate adjustments, and dropout regularization.

Early stopping, a regularization technique that halts training when the model's validation performance stops improving, was used to monitor validation loss. Validation loss is a metric that measures the discrepancy between the model's predictions and actual target values on the validation set using a predefined loss function. Training was terminated if no improvement occurred over 10 consecutive epochs.

The learning rate was initialized at a base value and systematically reduced by a factor of 0.75 every 5 epochs to facilitate stable convergence. Additionally, dropout regularization was applied to the fully connected layers with a rate of 0.5 to mitigate overfitting and enhance the model's generalization capabilities.

The models were trained using cross-entropy loss, a specific loss function used for classification tasks. It guides the training process by minimizing the difference between predictions and true labels.

Pixel intensity values were normalized to ensure consistency across the dataset. Each channel's mean and standard deviation were set to [0.5,0.5,0.5], which scales the pixel values to the range [−1,1]. This helps stabilize the training process and accelerates convergence.

Evaluations metrics for assessing the primary outcome:

The five models were assessed according to their precision, accuracy, recall and F1 score using the count of true positive (TP), false positive (FP), true negative (TN) and true false negative (FN) as shown in table 3.

III. RESULTS

After evaluating the performance of the three classification approaches—image-based classifier, site-enhanced classifier, and early site integration—the most compelling results were achieved with the site-enhanced classifier. Consequently, we selected the site-enhanced classifier method as our primary approach for further investigation, implementing V1, V2, and V3 variations.

Table 3: Evaluation metrics for the models' performances

Evaluation metrics	Definition	Formula	Indication
1. Precision	The accuracy of the positive predictions of the model.	Precision = TP / (TP+FP)	High precision values (reducing FP) allow the medical practitioners to avoid unnecessary interventions or referrals, especially crucial in diseases as oral cancer diagnosis and treatment.
2. Recall/ sensitivity/ true positive rate (TPR)	Quantifies how well a model can detect the lesions.	Recall/Sensitivity/TPR = TP/ (TP+FN)	Recall is essential in oral cancer screening, where higher values mean fewer missed cases (reducing FN) and validate model generalizability.
3. Accuracy	Measures the proportion of correctly predicted labels out of all labels	Accuracy = (TP+TN) / (TP+TN +FP+FN)	Accuracy offers a clear indication of the overall correctness of the model during testing.
4. F1 score (F-measure)	Provides a balanced evaluation of model's performance by assessing both precision and recall into a single metric.	F1-score= 2×Precision×Recall / (Precision +Recall)	It effectively balances the trade-off between identifying all relevant cases (recall) and ensuring the accuracy of predictions (precision).

Results of V1, V2 and V3 with the site-enhanced classifier approach in the five models is shown in table 4.

Regarding precision scores, the analysis revealed that GoogleNet and EfficientNet B4 achieved the highest precision scores of 0.81, with GoogleNet performing best under oversampling (V1) and EfficientNet B4 in the combined contralateral sides variant and without oversampling (V3). ResNet50 performed best without oversampling (V2), attaining a precision of 0.80, while ViT reached the precision of 0.79 with and without oversampling (V1 and V2).

In terms of accuracy, both GoogleNet and EfficientNet B4 reached the highest score of 0.82 with oversampling (V1). EfficientNet B4 maintained this performance in V3, while InceptionResNetV2 matched the same accuracy (0.82) without oversampling (V2). GoogleNet achieved the highest recall score of 0.81 under oversampling (V1), though its

performance slightly decreased in V3 (0.79). ViT and EfficientNet B4 followed closely, both attaining a recall of 0.79 in V1, with EfficientNet B4 maintaining the same score (0.79) in V3. InceptionResNetV2 demonstrated consistent results across V2 and V3 without oversampling, matching the aforementioned performance.

The F1-score was highest for GoogleNet and InceptionResNetV2 (0.81) in both V1 and V3. EfficientNet B4 closely followed with an F1-score of 0.80 in V3, reinforcing its robustness across different experimental setups.

The findings indicate that oversampling (V1) significantly improved model performance, particularly for GoogleNet, which consistently ranked among the top models in precision, accuracy, recall, and F1-score. InceptionResNetvr2 performed better in all evaluation metrics without oversampling. EfficientNet b4 showed similar results with and without oversampling, while ViT was the least consistent.

Table 4: Results of V1, V2 and V3 with the site-enhanced classifier approach in the five models.

Results of V1 with the site-enhanced classifier approach in the five models

Model name	Precision	Accuracy	Recall	F1 score
GoogleNet	0.81	0.82	0.81	0.81
ViT	0.79	0.81	0.79	0.78
EfficientNet B4	0.80	0.82	0.79	0.79
InceptionResNetvr2	0.80	0.81	0.78	0.81
Resnet 50	0.78	0.80	0.78	0.78

Results of V2 with the site-enhanced classifier approach in the five models

Model name	Precision	Accuracy	Recall	F1 score
GoogleNet	0.77	0.79	0.77	0.77
ViT	0.79	0.80	0.78	0.78
EfficientNet B4	0.80	0.81	0.78	0.79
InceptionResNetvr2	0.80	0.82	0.79	0.79
Resnet 50	0.80	0.80	0.76	0.78

Results of V3 with the site-enhanced classifier approach in the five models

Model name	Precision	Accuracy	Recall	F1 score
GoogleNet	0.80	0.81	0.79	0.81
ViT	0.77	0.78	0.74	0.74
EfficientNet B4	0.81	0.82	0.79	0.80
InceptionResNetvr2	0.79	0.81	0.79	0.81
Resnet 50	0.78	0.81	0.78	0.78

IV. DISCUSSION

For our study, we selected five pretrained models—GoogleNet, ViT, EfficientNet B4, InceptionResNetV2, and ResNet-50—based on their demonstrated efficacy in related research.

Tiryaki et al. achieved strong performance levels using ResNet-50 and GoogleNet for a 5-class tongue lesion classification task. ResNet-50 achieved 90.64% accuracy in detecting geographic tongue, while GoogleNet attained an F1-score of 82.01% for median rhomboid

glossitis classification and 81.66% in hairy tongue classification ***(Tiryaki et al., 2024)***. Lee et al. employed EfficientNet B0 on a small dataset of 1,810 tongue images, achieving an accuracy of 0.9167, precision of 0.9212, and recall of 0.9176, motivating our choice of EfficientNet B4 (Lee et al., 2022). Additionally, Rashid et al. reported 100% precision and recall using InceptionResNetV2 on a dataset of 517 intraoral images (augmented to 5,143) across 7 diseases, inspiring its application to our larger and more complex intraoral dataset ***(Rashid et al., 2024)***. These models were chosen for their proven performance and adaptability to medical image analysis tasks.

In our study, to address the common challenge of limited datasets in medical image analysis, three key approaches were implemented. First, the Dataset was expanded and diversified to include 5616 total original images, featuring over 30 distinct lesions. Second, On-the-fly data augmentation was employed in all of our experimental approaches to expand the size and diversity of the training set in deep learning. Third, to further mitigate dataset limitations, transfer learning was utilized. This allowed the models to leverage knowledge acquired from on large-scale datasets, enhancing their performance and generalization capabilities for the new classification task.

We used oversampling in V1 to compensate for class imbalance seen in sites that have fewer images in oral cancer such as the labial mucosa and gingiva as they are less susceptible to oral cancer. In V3, we assessed the impact of incorporating contralateral sides (e.g.: Upper and lower labial mucosa, right and left buccal mucosa, and right and left lateral border of tongue) to increase the representation of underrepresented areas by combining them together, which modified the 11 anatomical sites into 8 sites classification.

In cancer prognosis, high recall, low precision is generally prioritized because early detection of oral cancer is critical for effective treatment and improved patient outcomes. High recall minimizes false negatives, ensuring that fewer cancer cases are missed, which is essential given the severe consequences of delayed diagnosis. While low precision results in a higher number of false positives, leading to unnecessary follow-up tests or procedures, these are considered less harmful compared to the risks associated with undetected cancers. This approach aligns with the precautionary principle in medical diagnostics, where the benefits of identifying true positives early outweigh the inconveniences and costs of false positives, ultimately prioritizing patient safety and timely intervention.

Accordingly, oversampling along with on-the-fly augmentation provided better outcomes for cancer prognosis, as it achieves the highest recall while maintaining good precision. GoogleNet demonstrated superior performance with oversampling in V1, achieving both the highest recall (0.81) and precision (0.81) scores among all models. This dual advantage positions it as the optimal choice for clinical deployment, as it simultaneously minimizes missed cancer cases (false negatives) and reduces unnecessary interventions (false positives). While EfficientNet B4 matched GoogleNet's precision (0.81) in V3 (adding contralateral sides without oversampling), its recall remained lower (0.79 vs. 0.81). While further validation is needed, oversampling appears to maintain reasonable sensitivity and precision tradeoffs, potentially making it useful for early detection workflows.

The overall lower results compared to the aforementioned studies can be attributed to the higher complexity of our dataset, which includes larger image numbers, different pixel counts, greater variability in terms of lesion types, different intraoral anatomical sites, and risk categories. To investigate this hypothesis, we conducted a controlled experiment with InceptionResNetV2 using only 450 balanced images (without oversampling). While validation metrics appeared exceptionally high (99% accuracy, 98.7% recall), the subsequent severe performance drop during testing clearly demonstrated model overfitting. This contrast

underscores both the challenges of our comprehensive dataset and the risks of oversimplified training approaches.

V. CONCLUSION

The results reflect the inherent variability and complexity of the dataset, showcasing the challenges and opportunities in processing large-scale clinical oral image datasets. This marks a significant step forward in developing versatile models capable of addressing a wide spectrum of diseases.

Future efforts will focus on addressing the challenges posed by the high diversity of image dimensions and complexity of anatomical sites. Additionally, further research will investigate how data flows through different architectures to optimize processing efficiency and model performance.

Conflict of Interest:

The authors declare no conflict of interest.

Funding:

This research is self-funded and did not receive any public, commercial, or not-for-profit grants.

Ethics:

This research was approved by the ethical committee of the faculty of dentistry- Cairo University with ethical approval number 6-3-24.

Clinical trial registration:

The protocol for this study was registered on clinicaltrials.gov, under ID: NCT06325514

Credit statement:

Noran Ayman M. Ismael: Conceptualization, Data curation, Resources, Methodology, Visualization, Writing – original draft, Writing – review and editing.

Fat'heya M. Zahran: Data Curation, Methodology Supervision, Visualization, Writing - original draft, Writing - review & editing.

Yomna Safaa EL-Din: Conceptualization, Formal analysis, Methodology supervision, Writing - original draft, Writing - review & editing.

Noha Adel Azab: Resources, Data curation, Methodology Supervision, Visualization, Writing - original draft, Writing - review & editing.

Acknowledgments:

We extend our gratitude to **Engineer Waleed Khalid Alzamil,** for his technical efforts in developing and refining the AI model central to this work.

References

Abd Elrahman, S. M. and Abraham, A. (2013) 'A review of class imbalance problem', *Journal of Network and Innovative Computing,* 1, pp. 9-9.

Alexey, D. (2020) 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv preprint arXiv: 2010.11929*.

Alpaydin, E. (2020) *Introduction to machine learning.* MIT press.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I. and Jemal, A. (2024) 'Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: a cancer journal for clinicians,* 74(3), pp. 229-263.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L. and Haworth, A. (2021) 'A review of medical image data augmentation techniques for deep learning applications', *Journal of medical imaging and radiation oncology,* 65(5), pp. 545-563.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D. and Bray, F. (2015) 'Cancer

incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012', *International journal of cancer,* 136(5), pp. E359-E386.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. and Herrera, F. (2018) *Learning from imbalanced data sets.* Springer.

Ferreira, F., Silva, L. L. and Valente, M. T. 'Software engineering meets deep learning: a mapping study'. *Proceedings of the 36th annual ACM symposium on applied computing*, 1542-1549.

Ghaffar Nia, N., Kaplanoglu, E. and Nasab, A. (2023) 'Evaluation of artificial intelligence techniques in disease diagnosis and prediction', *Discover Artificial Intelligence,* 3(1), pp. 5.

Goceri, E. (2023) 'Medical image data augmentation: techniques, comparisons and interpretations', *Artificial Intelligence Review,* 56(11), pp. 12561-12605.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning.* MIT press Cambridge.

Han, D., Liu, Q. and Fan, W. (2018) 'A new image classification method using CNN transfer learning and web data augmentation', *Expert Systems with Applications,* 95, pp. 43-56.

He, K., Zhang, X., Ren, S. and Sun, J. 'Deep residual learning for image recognition'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z. and Azim, M. A. (2022) 'Transfer learning: a friendly introduction', *Journal of Big Data,* 9(1), pp. 102.

Ilhan, B., Lin, K., Guneri, P. and Wilder-Smith, P. (2020) 'Improving oral cancer outcomes with imaging and artificial intelligence', *Journal of dental research,* 99(3), pp. 241-248.

Jalata, I. K., Khan, R. and Nakarmi, U. (2024) 'Learning from Oversampling: A Systematic Exploitation of oversampling to address Data Scarcity issues in Deep Learning based Magnetic Resonance Image Reconstruction', *IEEE Access*.

Kandel, I. and Castelli, M. (2020) 'The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset', *ICT Express,* 6(4), pp. 312-315.

Kavyashree, C., Vimala, H. and Shreyas, J. (2024) 'A systematic review of artificial intelligence techniques for oral cancer detection', *Healthcare Analytics,* 5, pp. 100304.

Klang, E. (2018) 'Deep learning and medical imaging', *Journal of thoracic disease,* 10(3), pp. 1325.

Krishna, S. T. and Kalluri, H. K. (2019) 'Deep learning and transfer learning approaches for image classification', *International Journal of Recent Technology and Engineering (IJRTE),* 7(5S4), pp. 427-432.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. and Ferrante, E. (2020) 'Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis', *Proceedings of the National Academy of Sciences,* 117(23), pp. 12592-12594.

Lee, S.-J., Kwon, I.-J., Son, Y.-D., Kim, J.-H., Lee, M.-S., Kwon, D., Song, E.-S., Kim, B., Lee, J.-H. and Kim, H.-K. (2022) 'Early detection of tongue cancer using a convolutional neural network and evaluation of the effectiveness of EfficientNet'.

Neshat, M., Ahmed, M., Askari, H., Thilakaratne, M. and Mirjalili, S. (2024) 'Hybrid Inception Architecture with Residual Connection: Fine-tuned Inception-ResNet Deep Learning Model for Lung Inflammation Diagnosis from Chest Radiographs', *Procedia Computer Science,* 235, pp. 1841-1850.

Rajaraman, S., Ganesan, P. and Antani, S. (2022) 'Deep learning model calibration for improving performance in class-imbalanced

medical image classification tasks', *PloS one,* 17(1), pp. e0262838.

Rashid, J., Qaisar, B. S., Faheem, M., Akram, A., Amin, R. u. and Hamid, M. (2024) 'Mouth and oral disease classification using InceptionResNetV2 method', *Multimedia Tools and Applications,* 83(11), pp. 33903-33921.

Sengupta, N., Sarode, S. C., Sarode, G. S. and Ghone, U. (2022) 'Scarcity of publicly available oral cancer image datasets for machine learning research', *Oral Oncology,* 126, pp. 105737.

Shen, X., Tian, X., Liu, T., Xu, F. and Tao, D. (2017) 'Continuous dropout', *IEEE transactions on neural networks and learning systems,* 29(9), pp. 3926-3937.

Shield, K. D., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A. K., Bray, F. and Soerjomataram, I. (2017) 'The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012', *CA: a cancer journal for clinicians,* 67(1), pp. 51-64.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. 2014. Going deeper with convolutions. CoRR abs/1409.4842 (2014).

Tan, M. and Le, Q. 'Efficientnet: Rethinking model scaling for convolutional neural networks'. *International conference on machine learning*: PMLR, 6105-6114.

Tiryaki, B., Torenek-Agirman, K., Miloglu, O., Korkmaz, B., Ozbek, İ. Y. and Oral, E. A. (2024) 'Artificial intelligence in tongue diagnosis: classification of tongue lesions and normal tongue images using deep convolutional neural network', *BMC Medical Imaging,* 24(1), pp. 59.

Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., Jantana, P. and Vicharueang, S. (2022) 'AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer', *Plos one,* 17(8), pp. e0273508.

Welikala, R. A., Remagnino, P., Lim, J. H., Chan, C. S., Rajendran, S., Kallarakkal, T. G., Zain, R. B., Jayasinghe, R. D., Rimal, J. and Kerr, A. R. (2020) 'Automated detection and classification of oral lesions using deep learning for early detection of oral cancer', *IEEE Access,* 8, pp. 132677-132693.

Thermoplastic, Thermoplastic - Composite Materials,INTECH, 1st ed,Timisoara. Open Access Publisher **(2012)**. pp.25-48.

Salerno C, Pascale M, Contaldo M, Candida-associated denture stomatitis. Med Oral Patol Oral Cir Bucal. **(2011).**;16:139-143.

Menaka A. Abuzar, Suman Bellur, Nancy Duong, Billy B. Kim, Priscilla Lu, Nick Palfreyman, Dharshan Surendran, Vinh T. Tran Evaluating surface roughness of a polyamide denture base material in comparison with poly (methyl methacrylate) J Oral Sci **2010** 52, 577-581.