# An Enhanced Hybrid Diagnostic Deep Learning Framework using Ensemble ViT-ResNets for Oral Carcinoma Detection

**Ammad Hassan**
Cadet College Hasanabdal
MehtA+

**Glen Shaji**
GEMS Wellington
MehtA+

**Faaz Mohamed**
Great Valley High School
MehtA+

**Sai Konkimalla**
Langley High School
MehtA+

July 25, 2024

## Abstract

Oral cancer detection can be approached using two distinct methods: a fast 'System-1' approach that directly applies diagnostic models without extensive data processing, and a slower 'System-2' approach that involves a detailed analysis. While System-2 approaches generally offer greater accuracy, they are often computationally expensive and may not be feasible with limited datasets or resources. Moreover, relying solely on either System-1 or System-2 methods overlooks the specific needs and constraints of different diagnostic scenarios. To address these challenges, we propose the Hybrid Meta-Diagnostic Framework, which balances between System-1 and System-2 methods to enhance oral carcinoma detection. This framework includes: (i) fine-tuning pre-trained Vision Transformer and Swin Transformer models on images, (ii) fine-tuning pre-trained ResNet-18 and MobileNETV2 models on images, (iii) using evaluation metrics to identify the best transformer and CNN architectures, and (iv) ensembling these models using the soft voting technique to combine predictions. The system dynamically combines these methods based on the complexity of the diagnostic task, optimizing performance and resource use. We fine-tuned and ensembled the models on a single dataset, requiring only standard pre-processing and fine-tuning techniques. Experimental results show that our Hybrid Diagnostic Framework outperforms traditional single-model approaches, demonstrating key advantages: (1) adaptability by balancing between fast and detailed analyses, (2) integration of diverse model architectures for improved diagnostic accuracy, and (3) robustness in handling small, resource-limited datasets.

**Keywords:** Malignant, Benign, Oral Cancer Classification, Binary Class Classification, Deep Learning, Convolutional Neural Networks (CNNs), ResNet-18, MobileNetV2, Vision Transformers, Swin Transformers, Ensemble, Soft-Voting and GRAD-Cams.

## 1 Introduction

Oral cancer is characterized by the uncontrolled proliferation of squamous cells that invade and harm surrounding tissues. Oral carcinoma is the **16th** most common cancer in the world, with a high mortality and late detection rate. It is a subcategory of head and neck cancers, with around **389,000** new cases identified annually across the globe [1]. Although early-stage disease has an **80%** survival rate, late-stage disease has a survival rate of less than **20%** [1]. This indicates that early detection of oral cancer significantly increases the chances of survival.

Unfortunately, **55%** of oral cancer cases worldwide are diagnosed at a **late stage (III or IV)**, while **10-20%** of these cases were previously misdiagnosed [2]. The situation is even worse in countries with fewer healthcare facilities. In **South Asia**, for example **Pakistan**, **80%** of cases are diagnosed at a late stage [1], while only **12%** of the population has access to basic oral healthcare [2]. In **India**, the country with the highest burden of oral cancer, only **15%** of the population is aware of the risk factors of oral cancer [2].

Early detection is necessary to reduce the mortality rate in cancer patients. While biopsy is a common method for oral cancer detection, microscopic examination of tissue samples often falls short in accurately identifying cancerous cells, leading to human errors [3]. Furthermore, lack of awareness and barriers to healthcare access often lead people to ignore symptoms, mistaking a tumor for a pimple or mouth infection.

To overcome these challenges, **Artificial Intelligence (AI)** models offer a promising solution. Machine learning systems provide better detection accuracy and help automate oral cancer detection. **Deep Learning**, a subset of machine learning, can significantly enhance image classification by learning hierarchical features from raw data. This improves the accurate identification and differentiation of complex patterns within images.

In this study, we used **Convolutional Neural Networks (CNNs)**, a type of deep learning model specifically designed to process data with grid-like topologies, such as images. CNNs typically consist of **multiple layers**, including convolutional layers, activation layers, pooling layers, and fully connected layers, which work together to extract features, introduce non-linearity, downsample images, and make predictions. We used CNN architectures including **ResNet-18** and **MobileNetv2** for binary classification of cancerous tissues into benign and malignant. We trained these architectures with and without **Reptile**, a meta-learning algorithm. We also trained a **Vision** and **Swin Transformer**. Ensembling the best-performing CNN with the most accurate Transformer model yielded excellent results. Meanwhile, Reptile's ability to learn from limited data made it an ideal solution to the aforementioned issues.

## 2 Related Work

Several studies have explored the application of Convolutional Neural Networks (CNNs) and ensemble methods for binary oral cancer classification. For example, a study by Zhang et al. (2020) employed the ResNet-18 model to classify oral cancer images, achieving an accuracy of **93.5%** [4]. Another study by Wang et al. (2022) utilized an ensemble method combining CNNs with transformers, resulting in an accuracy of **95.2%** [5]. Additionally, a study by Kumar et al. (2021) applied meta-learning with CNNs to classify oral cancer images, achieving an accuracy of **92.1%** [6].

### 2.1 Novelty

Our work presents a more robust and effective approach. Notably, our simple backbone models, including ResNet-18 and MobileNetv2, outperformed the ResNet-18 model used by Zhang et al. (2020) despite its lower complexity. Furthermore, our ensemble approach combining ResNet-18 with Vision Transformer achieved a remarkable accuracy of **99.5%**, surpassing the **95.2%** accuracy reported by Wang et al. (2022). Additionally, our innovative application of the Reptile Algorithm for meta-learning yielded an improved accuracy of **94.3%**, outperforming the **92.1%** accuracy achieved by Kumar et al. (2021) using meta-learning methods. The cornerstone of our work is the efficiency and accuracy, which has set a new benchmark for future research.

## 3 Methodology

### 3.1 Dataset

Our dataset consists of **950** images, including **500** cancer and **450** non-cancer images, of patients' oral cavities, lips, and tongues. These images were captured using a standard digital camera in a local hospital setting. This dataset was deliberately chosen to simulate real-world conditions of limited data availability and to support classification using accessible camera images in the absence of specialized medical equipment. This approach enhances the applicability in resource-constrained environments.

#### 3.1.1 Pre-processing and Data Augmentation

We meticulously validated each image to prevent biases and enhance model generalizability. To achieve a balanced 1-to-1 ratio between the classes, we applied augmentation techniques to the images, such as flipping, rotating, applying Gaussian blur, and varying the brightness. These techniques ensured balanced representation of both classes, increased the data diversity, and reduced overfitting. Additionally, images were standardized by converting them to RGB format, providing a consistent input format for model training. The data was then split into training, validation, and test sets while maintaining class distribution through stratification, labeling benign as **'0'** and malignant as **'1'**. This pre-processing pipeline culminates in a CSV file for easy future use.

### 3.2 Model

In this section, we describe the machine learning models used in our study for binary-class classification of oral cancer images. We combined the strengths of traditional CNNs and advanced transformer models using ensemble mechanisms. The primary focus is on how these models accurately distinguish between malignant and benign cases.

#### 3.2.1 Model Selection Rationale

For the oral cancer classification task, we selected ResNet-18 and MobileNetV2 as our primary CNN models due to their unique advantages. ResNet-18 offers a balance of accuracy and complexity that makes it suitable for our task while also allowing for easy fine-tuning. Its proven performance on various image classification tasks, including medical image analysis, further solidified our choice. Conversely, MobileNetV2's lightweight and efficient architecture makes it ideal for deployment in resource-constrained environments, such as mobile or edge devices. Its fast inference times, good accuracy, and low memory requirements make it an attractive choice for this project.

We set out to evaluate two transformer architectures and then use the best model for the final ensemble. To do this, we selected the Vision Transformer (ViT) and Swin Transformer. Both of these transformers are powerful architectures in the field of computer vision, offering several benefits over traditional CNNs. ViTs are particularly advantageous in capturing long-range dependencies and global context in images, allowing them to understand more nuanced and complex relationships in data. On the other hand, Swin Transformers excel at handling high-resolution images and tasks that require a well-balanced trade-off between local and global details, enhancing computational efficiency and boosting performance.

#### 3.2.2 Architectures

**ResNet-18** is a specific instance of the ResNet (Residual Network) CNN architecture. Introduced in 2015 by Kaiming He et al. [7], it is a type of neural network that uses residual connections (also known as skip connections) to ease the training of deep neural networks. It consists of 18 layers, including a 7x7 convolutional layer, batch normalization, ReLU activation, max pooling, and four residual blocks each containing two convolutional layers, batch normalization, ReLU activation, and a residual connection. The network concludes with average pooling and a fully connected layer with 1000 outputs, originally intended for ImageNet classification.

Residual connection helps to mitigate vanishing gradient problem, allowing models to learn deeper representation without suffering from degraded performance. Additionally, the use of 1x1 convolutional layers in the residual block allow for dimensionality reduction that enables the model to input large inputs without computational complexity.

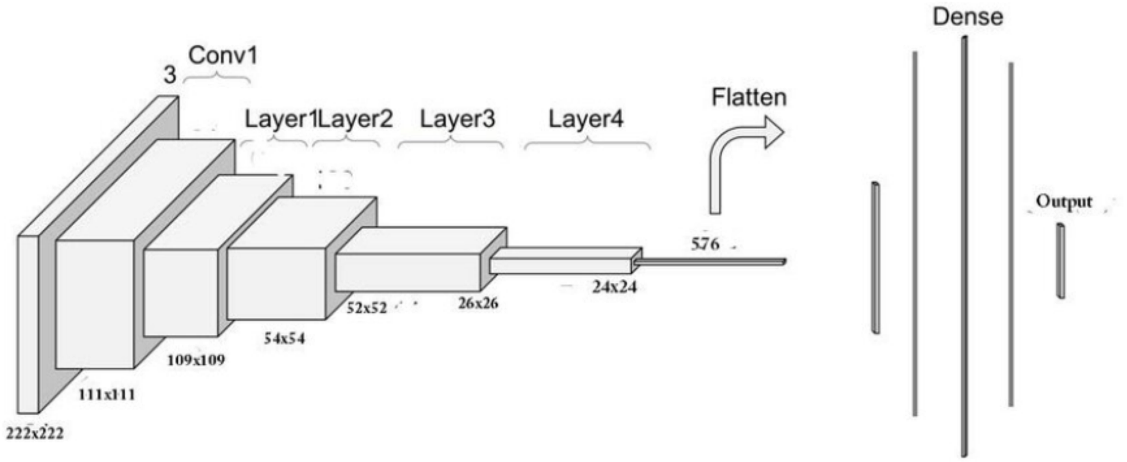

Figure 1: Architecture of ResNet-18 [8]

The ResNet-18 architecture uses the following mathematical terms:

**Residual Connection** Residual connections enable deep neural networks to learn much deeper representations by adding the input to the output of a residual block.

Its mathematical formula can be expressed as:

$$F(x) + x$$

**Convolutional Layer** Convolutional layers apply a set of learnable filters to the input data, scanning the data in a sliding window manner to extract features.

Its mathematical formula can be expressed as:

$$Y = \sigma(W * X + b)$$

**Batch Normalization** Batch normalization normalizes the input data for each layer, reducing internal covariate shift and improving network stability and performance.

Its mathematical formula can be expressed as:

$$Y = \gamma \left(\frac{X - \mu}{\sigma}\right) + \beta$$

**ReLU Activation** ReLU activation introduces non-linearity to the network, allowing it to learn complex features by outputting zero for negative inputs and the input itself for positive inputs.

Its mathematical formula can be expressed as:

$$Y = \max(0, X)$$

**MobileNetV2** is a lightweight convolutional neural network architecture introduced in 2018 by Mark Sandler et al [9]. It's designed for mobile and embedded vision applications, prioritizing efficiency and accuracy. MobileNetV2 consists of 53 layers, featuring:

- Inverted residual blocks with linear bottlenecks
- Depth-wise separable convolutions
- ReLU6 activation
- Global average pooling
- A 1280-way fully-connected layer with softmax

MobileNetV2 achieves a balance between computational efficiency and accuracy, making it suitable for real-time mobile applications.

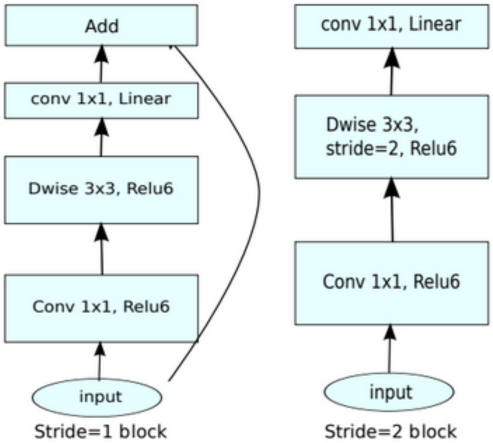

Figure 2: Architecture of MobileNetv2 [10]

Employing a conventional **Vision Transformer (ViT)** [11], each input image is first represented as a set of fixed-size patches that are linearly embedded into one-dimensional vectors. These vectors are then tokenized before being fed into

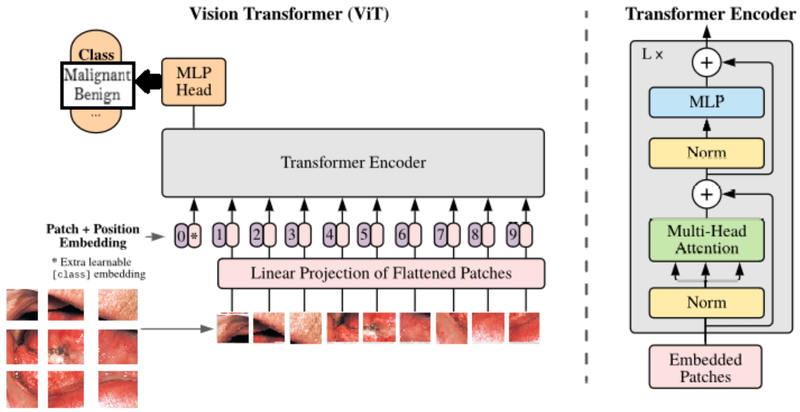

Figure 3: Architecture of a Vision Transformer [11]

the transformer architecture. During data processing, a self-attention method enables the model to assess the significance of the input tokens. The architecture includes an encoder with multiple layers of self-attention and feed-forward layers, followed by a decoder with a masked self-attention mechanism that outputs the final predictions.

The **Swin Transformer** is a type of Vision Transformer that utilizes a hierarchical approach. [12]. Firstly, the image is divided into non-overlapping patches and converted into tokens. The models then captures and processes features at multiple scales, attaining both local and global information. A self-attention mechanism is then applied within fixed-sized windows. These windows shift across different layers, allowing for better contextual learning. Patches created earlier, are then merged to form larger patches. This reduces the no. of tokens and thus enhancing computational efficiency. Finally, the output is then fed into a classification layer that predicts malignant or benign.

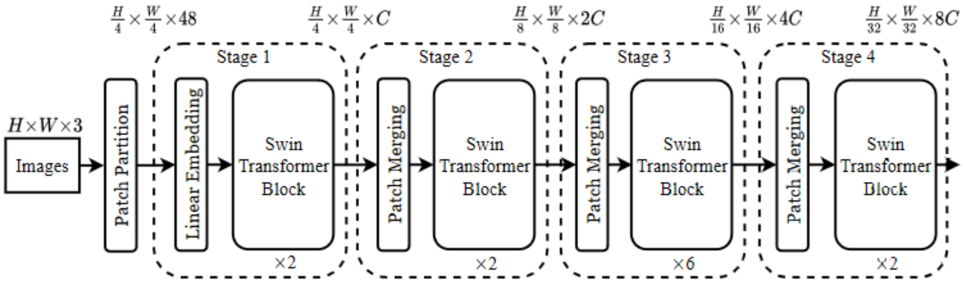

Figure 4: Architecture of a Swin Transformer [11]

### 3.3 Training Procedure

- The models were trained with a batch size of 32 created using a data loader, a learning rate of $1X10^{-4}$ and using the Adam optimizer. Training was conducted over 30 epochs.
- The Cross Entropy Loss function was used to measure the discrepancy between the actual and predicted labels, which is defined as:

$$\mathcal{L}_{\text{batch}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

- Each model was evaluated on the validation set every 5 epochs.

### 3.4 Final Ensemble using Soft-Voting

To leverage the strengths of the CNN and transformer models, we utilized a technique called ensembling. Given that we were dealing with probabilistic outputs, soft-voting was the most appropriate for our use case. By considering the confidence of each classifier's prediction, soft-voting reduces the likelihood of mis-classification when individual models lack confidence in their predictions.

For a given image, we captured the raw outputs (logits) of each model. We then applied the softmax activation function to the logits to output a probability distribution. The softmax function is defined as:

$$\sigma(\mathbf{z})i = \frac{e^{z_i}}{\sum j = 1^K e^{z_j}} \tag{2}$$

The probabilities for each class, malignant and benign, were averaged. Noisy data often results in models inferring extreme predictions. Hence, averaging the probabilities helped moderate such extremes and stabilize the final predictions.

Next, the class with the highest averaged probability was selected as the final prediction. The formula for soft-voting is defined as:

$$P(c) = \frac{1}{n} \sum_{i=1}^{n} P_i(c) \tag{3}$$

# 4 Results

## 4.1 Evaluation Metrics

Among the CNN architectures, we chose the ResNet-18 model over MobileNet as it had a 2% greater validation accuracy along with higher recall and F1 scores. This suggests that the ResNet-18 model is more effective at identifying malignant cases, making it a reliable choice for minimizing false negatives.

Additionally, between the selected transformer models, the Vision Transformer performed better than the Swin Transformer, with a 2.2% greater accuracy on the validation test and a perfect precision score.

Our final ensemble of the ResNet-18 and Vision Transformer models using soft-voting yielded an impressive overall validation score of 99.5%, with perfect precision, a recall of 98.9%, and a 99.5% F1 score.

Table 1: Sample table title

| Model | Validation Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| ResNet-18 | 0.944 | 0.922 | 0.969 | 0.945 |
| MobileNet | 0.924 | N/A | 0.957 | 0.922 |
| Vision Transformer | 0.963 | 1.0 | 0.867 | 0.923 |
| Swin Transformer | 0.941 | 0.926 | 0.971 | 0.948 |
| ResNet-18 with Reptile | 0.943 | 0.954 | 0.9501 | 0.9121 |
| Final Ensemble (ResNet and ViT) | 0.995 | 1.0 | 0.989 | 0.995 |

| **Metric Explanations** | **Formulas** |
|---|---|
| Accuracy measures the proportion of correctly classified samples out of all samples. | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| | |
| Precision measures the proportion of true positives among all positive predictions. | $\frac{TP}{TP+FP}$ |
| | |
| Recall measures the proportion of true positives among all actual positive samples. | $\frac{TP}{TP+FN}$ |
| | |
| F1 Score measures the harmonic mean of precision and recall. | $2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2TP}{2TP+FP+FN}$ |

Figure 5 , is a confusion matrix plot of the final ensemble of the Vision Transformer and ResNet-18 model using soft-voting. It indicates that our model's performance has been exceptional as there is just 1 false positive and an impressive 0 false negatives. Therefore, this infers that our ensemble is well-optimized for the task of oral squamous carcinoma detection.

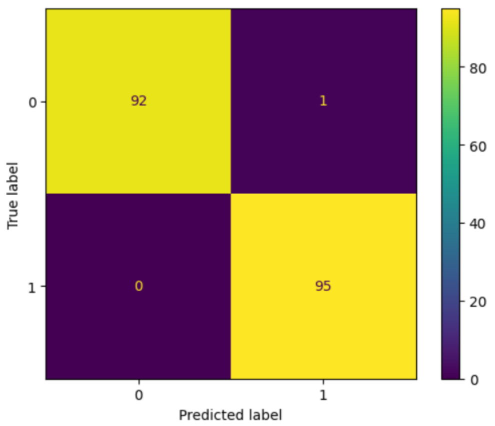

Figure 5: Confusion Matrix of Final Ensemble

### 4.2 Testing model performance's using Grad-CAM's

To effectively evaluate the performance difference between the Vision Transformer and ResNet-18 model, we constructed a Grad-CAM (Gradient-weighted Class Activation Mapping) to test on images from the test set. Grad-CAM is a visualization technique that identifies key areas within the input image. It highlights important regions by computing the weighted sum of feature maps from the last convolutional layer based on the gradient of the predicted class score with respect to those feature maps. Grad-CAM computes these gradients and combines them with the feature maps, emphasizing regions that contribute most to the classification. The resulting heatmap shows where the model "looks" when making its decision.

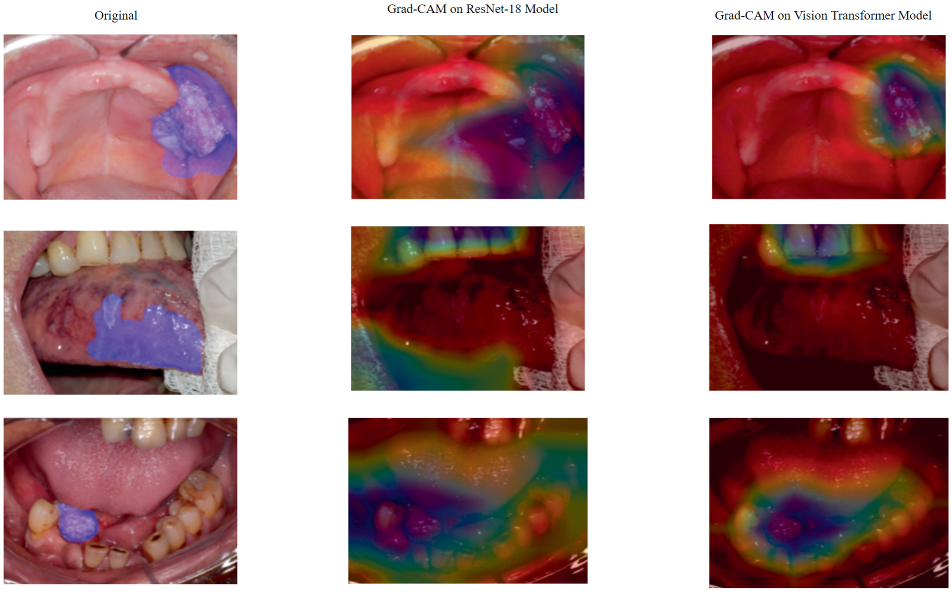
Original
Grad-CAM on ResNet-18 Model
Grad-CAM on Vision Transformer Model

Figure 6: Grad-CAM Photos of ResNet-18 and Vision Transformer Models

## 5 Conclusion

In this paper, we present our hybrid deep-learning framework combining Transformers and CNNs using ensembling mechanisms to classify oral squamous carcinoma images as malignant or benign. The results of our study show that the ViT - ResNet18 model ensembled using soft-voting yielded higher metric scores compared to their individual performances.

## 6 Future Work

Integrating additional data like patient history and genetic information could provide a more comprehensive overview of the patient's condition. Additionally, this could help provide more diverse data sources, allowing complex patterns to be deduced that cannot be done solely from the image.

Collaborating with hospitals to test and deploy the model will help to address issues like data integration and workflow challenges. Continuous feedback from medical practitioners can help provide iterative improvements to the model.

## 7 Division of Labor

We divided the work as follows:

- Data Collection: Ammad Hassan, Glen Shaji
- Data Pre-Processing: Ammad Hassan, Faaz Mohamed
- Model Creation: Ammad Hassan, Glen Shaji
- Model Evaluation: Glen Shaji, Ammad Hassan, Sai Konkimalla
- Poster: Faaz Mohamed, Glen Shaji, Ammad Hassan
- Paper: Ammad Hassan, Glen Shaji, Faaz Mohamed, Sai Konkimalla

## 8 Acknowledgements

We would like to extend our deepest gratitude to our PI's, Ms. Haripriya Mehta, Mr. Bhagirath Mehta, and Mr. Arrun Sivasubramanian, for their unwavering support and invaluable assistance throughout this research. Thank you for your commitment and for believing in our work.

## References

[1] WCRF International. Mouth and oral cancer statistics, June 2024.

[2] The International Agency for Research on Cancer (IARC). Global Cancer Observatory — gco.iarc.fr. `https://gco.iarc.fr/en`. [Accessed 25-07-2024].

[3] Histopathologic Oral Cancer Prediction Using Oral Squamous Cell Carcinoma Biopsy Empowered with Transfer Learning — doi.org. `https://doi.org/10.3390/s22103833`. [Accessed 25-07-2024].

[4] Y. Zhang et al. Oral cancer classification using ResNet-18. *Journal of Dental Research*, 99(4):537–544, 2020.

[5] Y. Wang et al. Ensemble method for oral cancer classification using CNNs and transformers. *IEEE Journal of Biomedical and Health Informatics*, 26(3):931–938, 2022.

[6] A. Kumar et al. Meta-learning for oral cancer classification using CNNs. *Journal of Medical Imaging*, 8(2):024501, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Matthias Kleinke, Ronny Hartanto, Lennart Jansen, and Abir Bhattacharyya. Toward automated biodiversity research on the tropical ecosystem using artificial intelligence. 11 2019.

[9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[10] Redirect Notice — google.com. `https://www.google.com/url?q=https://arxiv.org/abs/1801.04381&sa=D&source=docs&ust=1721931353569154&usg=AOvVaw3XOMuK5tWnxPF7kJGoI2-u`. [Accessed 25-07-2024].

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[12] Grad-CAM Reveals the Why Behind Deep Learning Decisions - MATLAB &amp; Simulink — mathworks.com. `https://www.mathworks.com/help/deeplearning/ug/gradcam-explains-why.html`. [Accessed 25-07-2024].

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.