Oral Health Project

Shaik Rehna Afroz (22B3932)
DH 307 R&D Project
IIT BOMBAY

Contents

1 Introduction

1.1 Problem Statement

- For detection of various oral pathologies like oral pre-cancerous lesions, caries, gingivitis, etc., it is required to identify various position, type of teeth, lesions etc, and identify any features required for better predictions.

1.2 Goal

- Fine-tune various state-of-the-art classification models on the given OPMD dataset Compare and evaluate the different models and identify the best one for identifying the pathology type of the oral cavity image

2 Literature Review

2.1 Screening of oral potentially malignant disorders and oral cancer using deep learning models

Karishma Madhusudan Desai1,2, Pragya Singh 2, Mahima Smriti2, VivekTalwar3, Manav Chaudhary2, George Paul2, Subhas Chandra Kolli2, Parisa Sai Raghava2, Golla Vamshi Krishna2, C. V. Jawahar3, P. K.Vinod 4,5, Varma Konala2,6 & *Ramanathan Sethuraman6 Published: 23 May 2025*

2.1.1 Background & Motivation

- Oral cancer rates remain high in low- and middle-income countries, largely due to late diagnoses
- Oral potentially malignant disorders (OPMDs) often precede cancer but are frequently undetected
- There is a need for automated, easy-to-use, point-of-care screening tools, particularly in resource-limited settings

2.1.2 Objective

- Evaluate AI-driven screening tools for detecting OPMDs and oral cancers using images captured via smartphones
- Investigate deployment in two scenarios: cloud-based web apps and native mobile (on-device) apps

2.1.3 Dataset & Annotation

- Utilized Grace dataset of 518 intraoral images, annotated as ”suspicious” or ”non-suspicious”
- Standardized protocols were followed for lesion annotation and classification

2.1.4 Models Evaluated

- Two deep learning architectures were evaluated:
  DenseNet-201: A powerful pre-trained convolutional neural network (20 million parameters)
  FixCaps: A lightweight capsule network variant (0.83 million parameters), trained from scratch
- Purpose: Compare a robust but compute-intensive model (DenseNet-201, 20 million parameters) with a leaner alternative (FixCaps, 0.83 million parameters) suitable for running on smartphones

2.1.5 Performance Metrics

- DenseNet-201:
  F1 score: 0.875
  AUC (Area Under Curve): 0.97
  Overall accuracy: 88.6
- FixCaps:
  F1 score: 0.828
  AUC: 0.93
  Accuracy: 83.8

2.1.6 Deployment Potential

- DenseNet-201 shows high performance and is implementable via web/cloud applications.
- FixCaps, with its significantly smaller size, enables deployment as a native smartphone application, highlighting practical usability in low-resource settings
- Both models offer effective screening tools for OPMDs and oral cancer using smartphone images.
- DenseNet-201 achieves slightly better accuracy and discrimination, while FixCaps offers efficiency and better suitability for offline/mobile use.
- The study underscores the feasibility of deploying AI-based screening tools in real-world, low-infrastructure environments.

2.2 Classification of Oral Potentially Malignant Disorders Using Multimodal Feature Integration

Buddhadev Goswami†1 Susmit Neogi†2 Saurabh Nagar3 Nirmal Punjabi1 Ravindra Gudi4

- This paper proposed a novel hybrid framework for early detection of Oral Potentially Malignant Disorders (OPMDs)

2.2.1 Methods

- *Multimodal Feature Integration:*This paper proposes a framework that merges standard RGB imaging with hyperspectral imaging (HSI) features. To enrich RGB data without the need for expensive equipment, a Multi-stage Spectral-wise Transformer (MST++) is employed for spectral reconstruction—even when only RGB input is available

- *Feature Engineering:* A diverse set of features is extracted to create a robust representation: Hu Moments (shape descriptors) Haar texture features SIFT descriptors combined with a Bag-of-Visual-Words model RGB statistical features Hyperspectral spectral features via MST++ This comprehensive feature fusion helps capture both spatial, textural, and spectral characteristics of lesions

- *Classification Model:* The paper uses an optimized Random Forest classifier. This choice is well-suited for handling high-dimensional, multimodal feature vectors, especially when datasets are limited or imbalanced

2.2.2 Advantages & Context

- *Data-Efficient & Practical:* By reconstructing spectral information from RGB data through MST++, the framework avoids reliance on costly hyperspectral imaging hardware—making it more viable in resource-constrained clinical settings

- *Adaptable to Limited Datasets:* Instead of deep learning models that require large annotated datasets, this approach relies more on handcrafted features and classical ML (Random Forests), which can better handle small or imbalanced datasets typical in medical imaging

2.3 ResNet

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

2.3.1 Motivation

- Deeper neural networks usually perform better (more representational power)

- But as depth increases, training becomes difficult:
  Vanishing/exploding gradients → optimization problem.
  Degradation problem → deeper models sometimes have higher training error than shallower ones.

2.3.2 Key Idea

- Then output becomes: y=F(x)+x

- The shortcut (identity connection) allows gradients to flow easily backward, solving vanishing gradient issues.

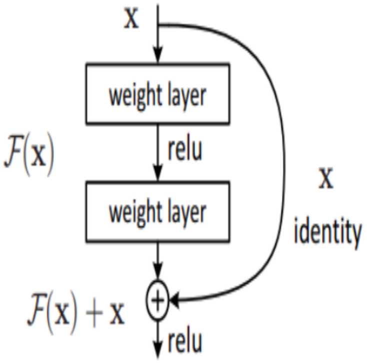
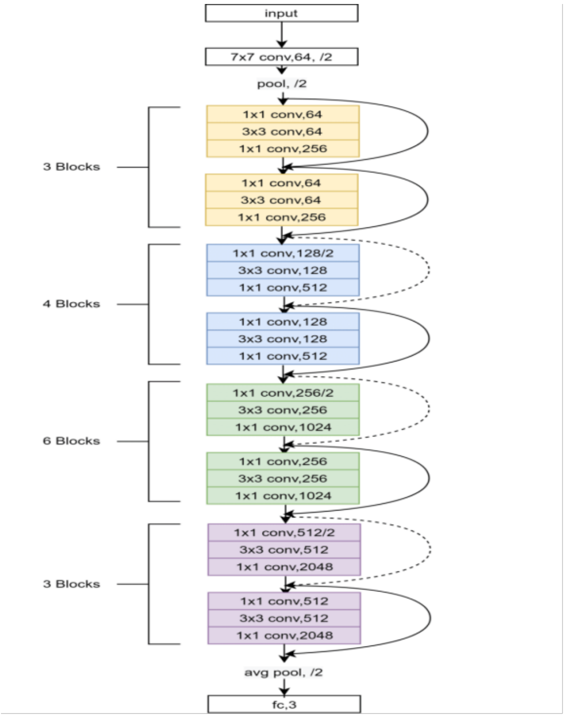

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning: a building block.



2.4 Densely Connected Convolutional Networks

Gao Huang, Zhuang Liu, Laurens van der Maaten,Kilian Q. Weinberger

- Dense Block: group of densely connected layers
- Transition Layer: reduces feature map size and number of channels (using 1×1 conv + pooling).
- Growth Rate (k): number of new feature maps each layer adds. Keeps network compact.

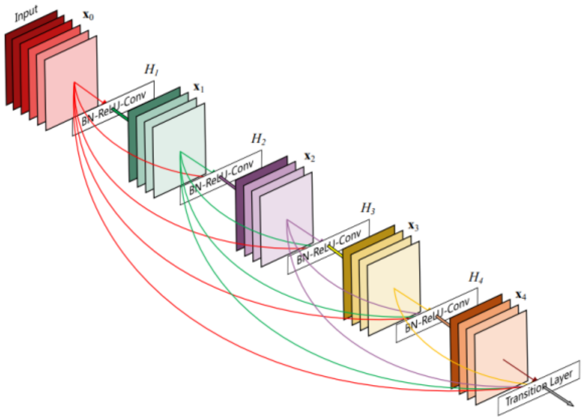


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

2.5 ResNet vs DenseNet

Feature Reuse

- **ResNet**: Features are passed forward through additive identity shortcuts. Earlier representations are not concatenated but influence deeper layers implicitly through residual connections. This helps optimization but does not enforce explicit reuse of past feature maps.
- **DenseNet**: Every layer receives *all* previous feature maps by direct concatenation. This creates explicit feature reuse, rich representations, and strong gradient flow, encouraging the network to build upon rather than relearn features.

Parameters and Efficiency

- **ResNet**: Uses more parameters because each layer independently learns new feature maps. Feature dimensionality remains stable since maps are not concatenated. Example: ResNet-50 has approximately 25M parameters.
- **DenseNet**: Uses fewer parameters due to extensive feature reuse. The growth rate k determines how many new feature maps each layer adds. Example: DenseNet-121 has about 8M parameters, far smaller than ResNet-50 for comparable accuracy.

Gradient Flow

- **ResNet**: Gradients propagate efficiently via residual shortcuts, mitigating vanishing gradients.
- **DenseNet**: Enables even stronger gradient flow because each layer has multiple direct gradient paths from all subsequent layers.

Memory and Computation

- **ResNet**: More memory efficient, with computation scaling linearly with depth.
- **DenseNet**: Requires storing feature maps from all preceding layers, increasing memory usage despite fewer parameters. DenseNet-201 has roughly 20M parameters, much fewer than a similarly deep ResNet-200 (64M parameters).

3 Dataset

- Piyarathne et al OPMD dataset
  3000 images
  Size : 1200 x 1600

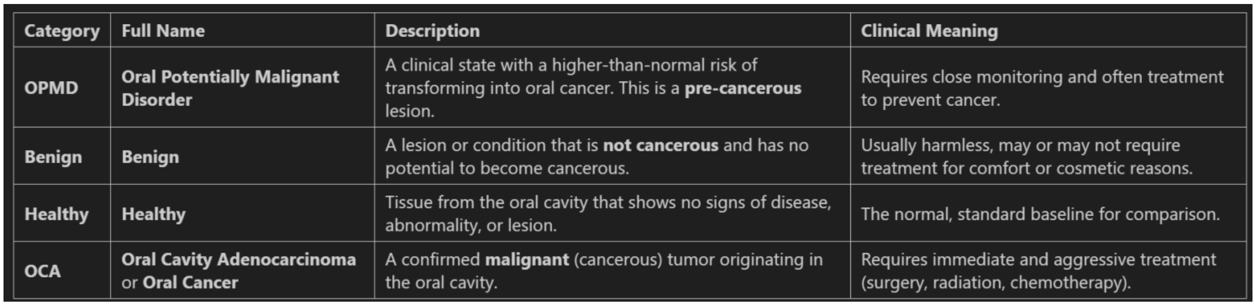
4 Data Available

- Data Available:
  IMAGE DIR Images (3000 images)
  ANNOTATION FILE (Annotation.json)
  IMAGEWISE METADATA FILE(imagewise data.xlsx)
  PATIENTWISE METADATA FILE (patientwise data.xlsx)

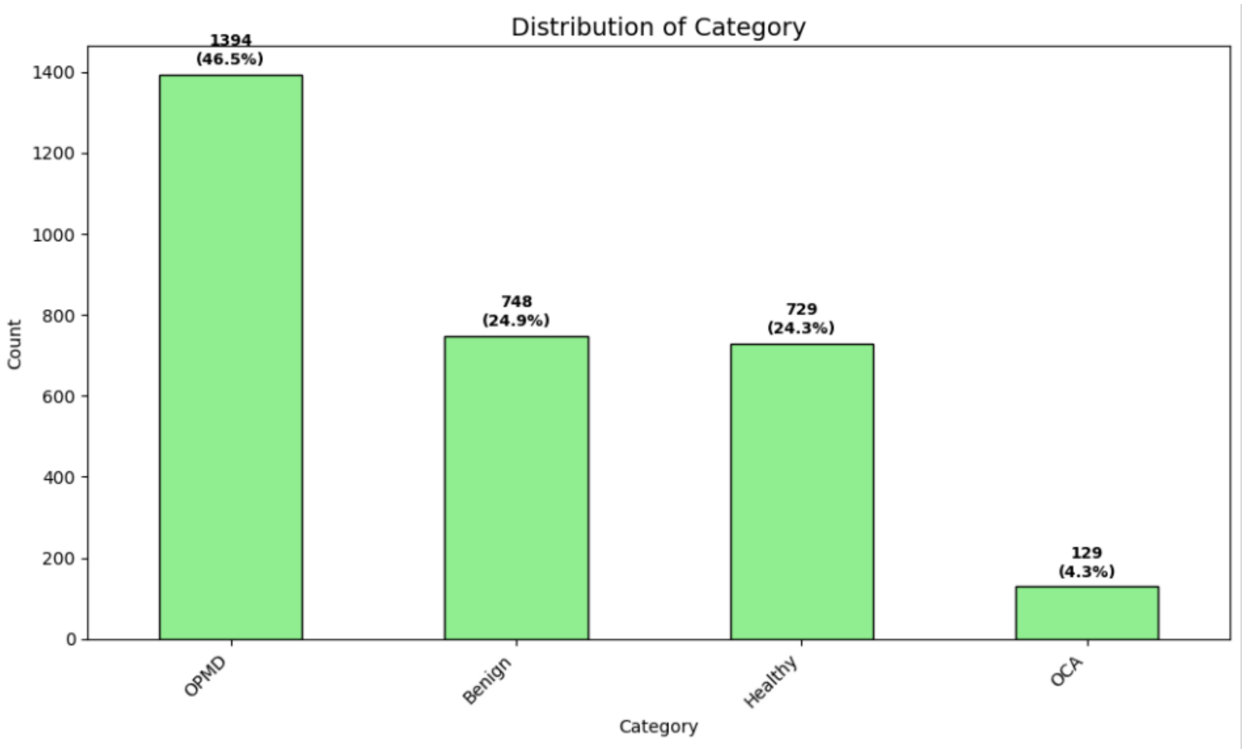
5 Categories

The dataset comprises four categories:

- OPMD
- Benign
- Healthy
- OCA



Category	Full Name	Description	Clinical Meaning
OPMD	**Oral Potentially Malignant Disorder**	A clinical state with a higher-than-normal risk of transforming into oral cancer. This is a **pre-cancerous** lesion.	Requires close monitoring and often treatment to prevent cancer.
Benign	**Benign**	A lesion or condition that is **not cancerous** and has no potential to become cancerous.	Usually harmless, may or may not require treatment for comfort or cosmetic reasons.
Healthy	**Healthy**	Tissue from the oral cavity that shows no signs of disease, abnormality, or lesion.	The normal, standard baseline for comparison.
OCA	**Oral Cavity Adenocarcinoma** or **Oral Cancer**	A confirmed **malignant** (cancerous) tumor originating in the oral cavity.	Requires immediate and aggressive treatment (surgery, radiation, chemotherapy).



6 Methodology

The following steps were followed during training:

- Input Image
- Preprocessing
- Classifier (DL Model)

6.1 Preprocessing

The preprocessing steps include:

- Cropping
- Resizing (224×224)
- Data Augmentation:
  - **Geometric**: Horizontal flips, rotations
  - **Photometric**: Color jitter (brightness, contrast, saturation, hue)

6.2 Classification Models

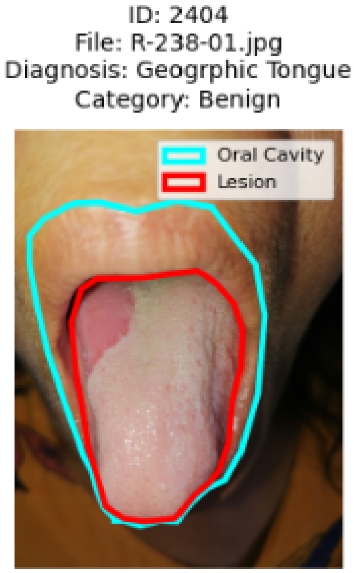
The following deep learning models were used for classification:

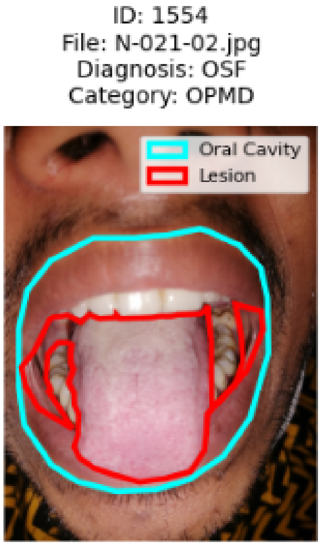
- ResNet18
- ResNet50
- DenseNet121
- EfficientNet-B0
- EfficientNet-B3

6.2.1 Type of Classification

This is a Mutli class classification problem. 3 classes have been chosen:

- Healthy
- OPMD
- OCA



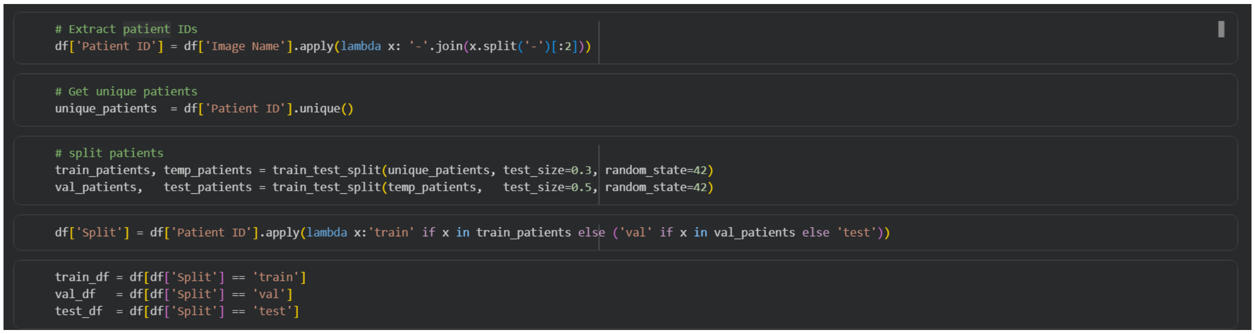


6.3 Removing Benign, OSF Images

- If we observe the annotations for Benign category, the entire tongue region or the most of the oral cavity region is considered as lesion.This can confuse the model about the healthy regions.Hence excluding the Benign images from the data lead to a better trained model
- OSF is a condition where the patient finds it difficult to completely open his mouth due to stiffened muscles. So the lesions cannot be marked properly. Hence removing these OSF images can lead to a better model
- So Benign and OSF images are removed from the image data
- The no of images after removing Benign, OSF is 1568
- Count of each Category in the dataset:
  Healthy 729 (46.49 %)
  OPMD 710 (45.28 %)
  OCA 129 (8.227 %)
- The dataset is imbalanced with Imbalance Ratio around 6:1
- To address this ”Weighted Cross Entropy Loss” is implemented

7 Splitting the Data

- The data has to be split in such a way that, all the images of a patient should only be in one of train, val, test. No data leakage should happen at patient level. This is crucial for medical imaging because otherwise the model can ”cheat” by seeing very similar images in training and validation/testing
- For this all the unique patient IDs are obtained and then these are split into train, val, test All the images related to certain ID stays in one of the split



```
# Extract patient IDs
df['Patient ID'] = df['Image Name'].apply(lambda x: '-'.join(x.split('-')[:2]))

# Get unique patients
unique_patients  = df['Patient ID'].unique()

# split patients
train_patients, temp_patients = train_test_split(unique_patients, test_size=0.3, random_state=42)
val_patients,   test_patients = train_test_split(temp_patients,   test_size=0.5, random_state=42)

df['Split'] = df['Patient ID'].apply(lambda x:'train' if x in train_patients else ('val' if x in val_patients else 'test'))

train_df = df[df['Split'] == 'train']
val_df   = df[df['Split'] == 'val']
test_df  = df[df['Split'] == 'test']
```

7.1 Criterion, Optimizer & Scheduler

- Criterion: Weighted Cross Entropy Loss
- Optimizer: Adam / AdamW
- Scheduler: CosineAnnealingLR

8 Results

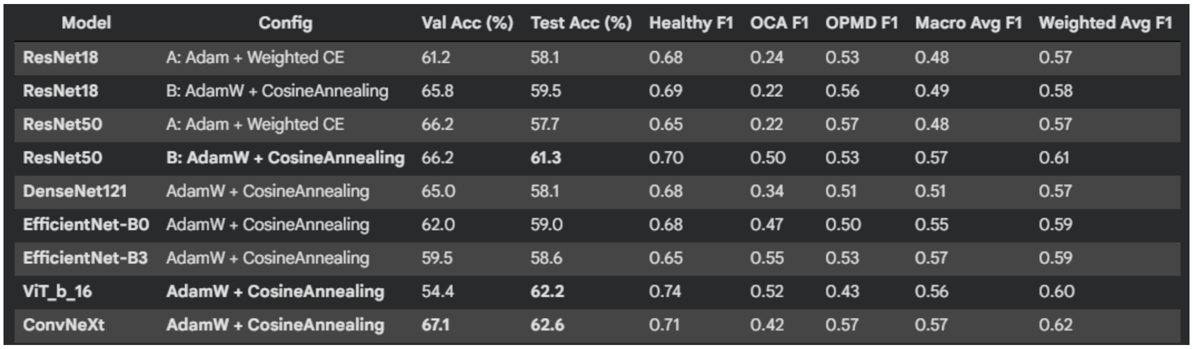
8.1 Experiment 1

Input

- Raw images resized to 224×224.

Train Set Data Transformations

- RandomHorizontalFlip($p = 0.5$)
- RandomRotation(10°)
- ColorJitter(brightness=0.1, contrast=0.1, saturation=0.1, hue=0.05)
- ToTensor()
- Normalize(mean=[0.5088, 0.3543, 0.3055], std=[0.2845, 0.2283, 0.2253])



Model	Config	Val Acc (%)	Test Acc (%)	Healthy F1	OCA F1	OPMD F1	Macro Avg F1	Weighted Avg F1
ResNet18	A: Adam + Weighted CE	61.2	58.1	0.68	0.24	0.53	0.48	0.57
ResNet18	B: AdamW + CosineAnnealing	65.8	59.5	0.69	0.22	0.56	0.49	0.58
ResNet50	A: Adam + Weighted CE	66.2	57.7	0.65	0.22	0.57	0.48	0.57
ResNet50	**B: AdamW + CosineAnnealing**	66.2	**61.3**	0.70	0.50	0.53	0.57	0.61
DenseNet121	AdamW + CosineAnnealing	65.0	58.1	0.68	0.34	0.51	0.51	0.57
EfficientNet-B0	AdamW + CosineAnnealing	62.0	59.0	0.68	0.47	0.50	0.55	0.59
EfficientNet-B3	AdamW + CosineAnnealing	59.5	58.6	0.65	0.55	0.53	0.57	0.59
ViT_b_16	**AdamW + CosineAnnealing**	54.4	**62.2**	0.74	0.52	0.43	0.56	0.60
ConvNeXt	**AdamW + CosineAnnealing**	**67.1**	**62.6**	0.71	0.42	0.57	0.57	0.62

Model Performance Summary

1. ConvNeXt – AdamW + CosineAnnealingLR

- Test Accuracy: 62.6%
- Highest Validation Accuracy: 67.1%
- Best Weighted F1: 0.62
- Most balanced and reliable performance

2. ViT_b_16 – AdamW + CosineAnnealingLR

- Test Accuracy: 62.2%
- Validation Accuracy: 54.4%
- Best Healthy class detection (F1 = 0.74)

3. ResNet50 – AdamW + CosineAnnealingLR

- Test Accuracy: 61.3%
- Validation Accuracy: 66.2%
- Good OCA detection (F1 = 0.50)

Training Configuration Impact

- AdamW + CosineAnnealingLR generally outperforms basic Adam.
- ResNet50 shows significant improvement with better optimization.

Class-Wise Performance Analysis

Healthy Class

- Best: ViT_b_16 (F1 = 0.74)

OCA (Critical Minority Class)

- Best: EfficientNet-B3 (F1 = 0.55)

OPMD Class

- Best: ResNet50-A & ConvNeXt (F1 = 0.57)

Validation vs Test Consistency

- **Most Consistent:** ConvNeXt (67.1% → 62.6%)
- **Overfitting:** ViT_b_16 (54.4% → 62.2%)

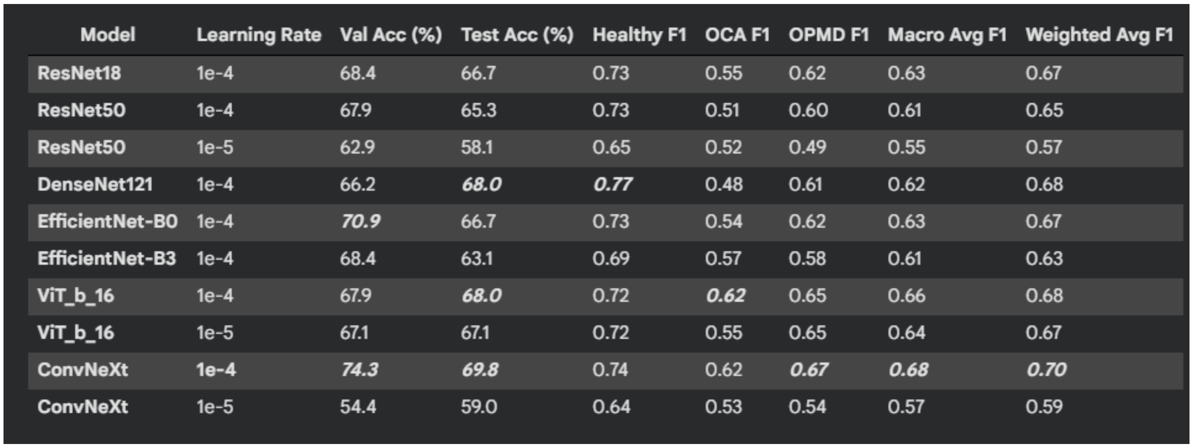
8.2 Experiment 2

Input

- Oral cavity regions tightly cropped and resized to 224×224.

Train Set Data Transformations

- RandomHorizontalFlip($p = 0.5$)
- RandomRotation(10°)
- ColorJitter(brightness=0.1, contrast=0.1, saturation=0.1, hue=0.05)
- ToTensor()
- Normalize(mean=[0.5091, 0.3415, 0.3140], std=[0.3110, 0.2445, 0.2411])



Model	Learning Rate	Val Acc (%)	Test Acc (%)	Healthy F1	OCA F1	OPMD F1	Macro Avg F1	Weighted Avg F1
ResNet18	1e-4	68.4	66.7	0.73	0.55	0.62	0.63	0.67
ResNet50	1e-4	67.9	65.3	0.73	0.51	0.60	0.61	0.65
ResNet50	1e-5	62.9	58.1	0.65	0.52	0.49	0.55	0.57
DenseNet121	1e-4	66.2	***68.0***	***0.77***	0.48	0.61	0.62	0.68
EfficientNet-B0	1e-4	***70.9***	66.7	0.73	0.54	0.62	0.63	0.67
EfficientNet-B3	1e-4	68.4	63.1	0.69	0.57	0.58	0.61	0.63
ViT_b_16	1e-4	67.9	***68.0***	0.72	***0.62***	0.65	0.66	0.68
ViT_b_16	1e-5	67.1	67.1	0.72	0.55	0.65	0.64	0.67
ConvNeXt	**1e-4**	***74.3***	***69.8***	0.74	0.62	***0.67***	***0.68***	***0.70***
ConvNeXt	1e-5	54.4	59.0	0.64	0.53	0.54	0.57	0.59

Training Configuration

- LR: 1×10^{-4} or 1×10^{-5}
- Loss: Weighted Cross-Entropy
- Optimizer: AdamW

Model Results

1. ConvNeXt (LR = 1×10^{-4})

- Test Accuracy: 69.8%

2. ViT_b_16 (LR = 1×10^{-4})

- Test Accuracy: 68.0%

3. DenseNet121 (LR = 1×10^{-4})

- Test Accuracy: 68.0%

4. ResNet18 (LR = 1×10^{-4})

- Test Accuracy: 66.7%

Learning Rate Optimization

Optimal Learning Rate

- 1×10^{-4} consistently performs best.

Model-wise Comparison

- ViT: $1e{-}4 > 1e{-}5$
- ConvNeXt: $1e{-}4 > 1e{-}5$

Class-Wise Performance

Healthy

- Best: DenseNet121 (F1 = 0.77)

OCA

- Best: ViT_b_16, ConvNeXt (F1 = 0.62)

OPMD

- Best: ConvNeXt (F1 = 0.67)

8.3 Experiment 3

Dataset

- Piyarathne et al. (2024) Zenodo OPMD dataset(3000 images) + extra OPMD images(874 Images)

Input

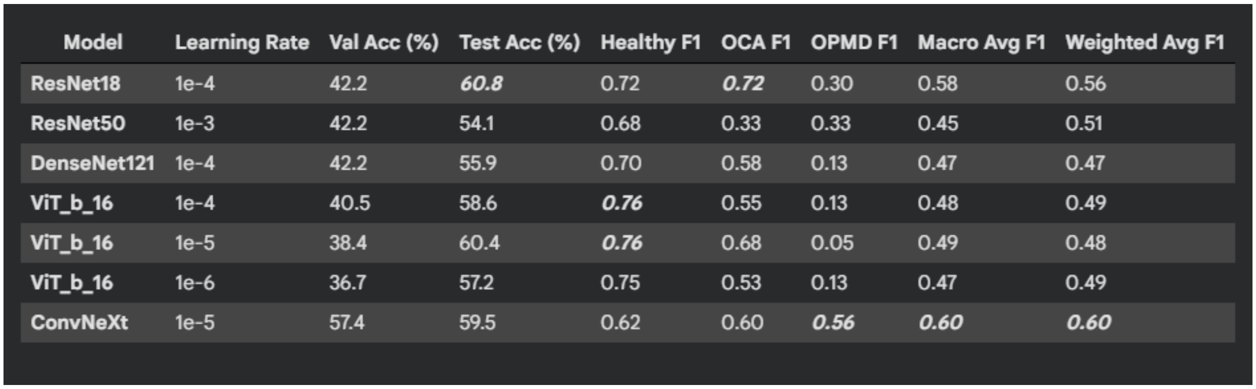
- Oral cavity crops (224×224)

Normalization Statistics

- Mean: [0.6246, 0.4630, 0.4216]
- Standard Deviation: [0.2522, 0.2333, 0.2270]

Train Set Data Transformations

- RandomHorizontalFlip, RandomRotation



Model	Learning Rate	Val Acc (%)	Test Acc (%)	Healthy F1	OCA F1	OPMD F1	Macro Avg F1	Weighted Avg F1
ResNet18	1e-4	42.2	***60.8***	0.72	***0.72***	0.30	0.58	0.56
ResNet50	1e-3	42.2	54.1	0.68	0.33	0.33	0.45	0.51
DenseNet121	1e-4	42.2	55.9	0.70	0.58	0.13	0.47	0.47
ViT_b_16	1e-4	40.5	58.6	***0.76***	0.55	0.13	0.48	0.49
ViT_b_16	1e-5	38.4	60.4	***0.76***	0.68	0.05	0.49	0.48
ViT_b_16	1e-6	36.7	57.2	0.75	0.53	0.13	0.47	0.49
ConvNeXt	1e-5	57.4	59.5	0.62	0.60	***0.56***	***0.60***	***0.60***

Model Performance

1. ResNet18 (LR = $1e{-}4$)

- Test Accuracy: 60.8%

2. ConvNeXt (LR = $1e{-}5$)

- Test Accuracy: 59.5%

3. ViT_b_16 (LR = $1e{-}5$)

- Test Accuracy: 60.4%

8.4 Experiment 4

Dataset

- Piyarathne et al. (2024) OPMD dataset(3000 images) + extra OPMD images(874 Images)

Input

- Cropped oral cavity images

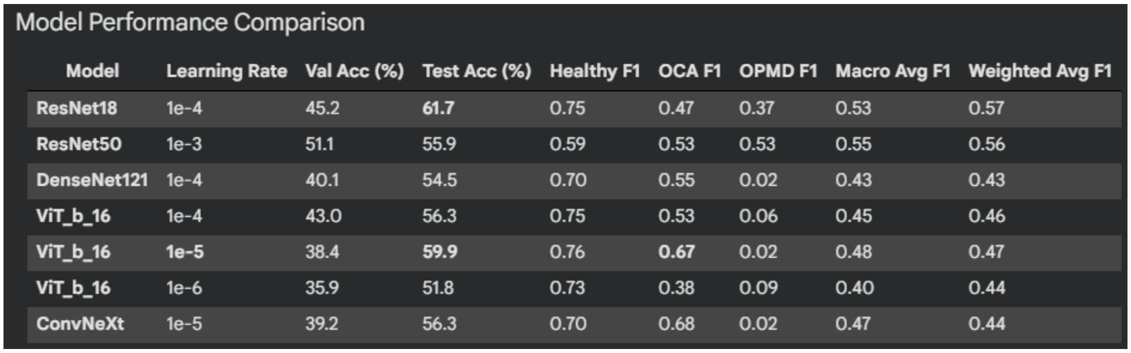
Normalization Statistics

- ImageNet mean/std
- imagenet mean = [0.485, 0.456, 0.406]
- imagenet std = [0.229, 0.224, 0.225]

Model Performance

1. ResNet18

- Test Accuracy: 61.7%



Model Performance Comparison

Model	Learning Rate	Val Acc (%)	Test Acc (%)	Healthy F1	OCA F1	OPMD F1	Macro Avg F1	Weighted Avg F1
ResNet18	1e-4	45.2	**61.7**	0.75	0.47	0.37	0.53	0.57
ResNet50	1e-3	51.1	55.9	0.59	0.53	0.53	0.55	0.56
DenseNet121	1e-4	40.1	54.5	0.70	0.55	0.02	0.43	0.43
ViT_b_16	1e-4	43.0	56.3	0.75	0.53	0.06	0.45	0.46
ViT_b_16	**1e-5**	38.4	**59.9**	0.76	**0.67**	0.02	0.48	0.47
ViT_b_16	1e-6	35.9	51.8	0.73	0.38	0.09	0.40	0.44
ConvNeXt	1e-5	39.2	56.3	0.70	0.68	0.02	0.47	0.44

2. ViT_b_16

- Test Accuracy: 59.9%

3. ResNet50

- Test Accuracy: 55.9%

9 Observations

- AdamW optimizer gave higher accuracy than Adam
- Training on cropped oral cavity images gave better performance than training on raw images.
- Adding extra OPMD images(874 Images) from other dataset to the Piyarathne et al. (2024) OPMD dataset(3000 images) degraded the performance.
- Using the mean and standard deviation obtained from the training data gives better performance than using Imagenet statistics

10 Conclusion

- Training on tightly cropped oral cavity regions from the Piyarathne et al. (2024) Zenodo OPMD dataset provides the highest validation accuracy, test accuracy, and F1 scores across experiments.
- Among all evaluated models, **ConvNeXt** achieved the best overall performance, particularly when trained with:
- Learning rate: 1×10^{-4}
- Optimizer: AdamW
- Scheduler: CosineAnnealingLR

This configuration offers strong generalization, balanced per-class detection, and superior performance on clinically important classes such as OPMD and OCA.

11 Future Work

- **Enhanced Cropping via Segmentation:** Integrate advanced segmentation models (U-Net, Mask R-CNN) to generate more accurate and standardized oral cavity crops, reducing noise introduced by imperfect bounding regions.
- **Clinically Interpretable AI Pipeline:** Incorporate interpretability tools such as Grad-CAM++, Score-CAM, and uncertainty maps to highlight lesion-relevant regions, improving model transparency for OPMD/OCA diagnosis in clinical settings.
- **Model Deployment and Optimization:** Apply model compression techniques—quantization, pruning, and knowledge distillation—and convert models to mobile-friendly formats for real-time inference on portable oral screening devices.