TRANSFER LEARNING FOR MULTI-LABEL RETINAL DISEASE CLASSIFICATION

Safi Shah
syshah25@student.oulu.fi
ITEE, University of Oulu

ABSTRACT

This work investigates transfer learning strategies for detecting Diabetic Retinopathy, Glaucoma, and Age-related Macular Degeneration using the ODIR dataset. Pretrained EfficientNet and ResNet18 models are evaluated under different fine-tuning schemes. Additionally, loss functions designed for class imbalance and attention mechanisms are explored to further improve performance. Experimental results demonstrate the impact of fine-tuning depth, loss reweighting, and attention modules on classification accuracy, measured using precision, recall, and F-score metrics. Finally, a custom ensemble solution is trained using deep learning models.

***Index Terms*—** Transfer learning, multi-label classification, medical imaging, deep learning

1. INTRODUCTION

In this project, I focused on multi-label classification of three major retinal diseases: Diabetic Retinopathy (DR), Glaucoma (G), and Age-related Macular Degeneration (AMD). Using the ODIR dataset, I evaluated different transfer learning strategies, loss functions addressing class imbalance, attention mechanisms to improve classification performance, and custom ensemble solutions to improve performance.

2. METHODS

2.1. Dataset Description and Augmentation

The dataset contains retinal fundus images labeled for diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD). The training set has 800 images, with DR being most common (64.6% positive), followed by glaucoma (20.4%) and AMD (17.8%). Validation and offsite test sets each have 200 images, showing similar distributions. Class imbalance is addressed using weighted binary cross-entropy with weights 0.5474 (DR), 3.9080 (glaucoma), and 4.6338 (AMD).

Training uses on-the-fly **augmentation** to expand variability, including rotations, flips, crops, color jitter, affine and shear transforms, grayscale, and elastic deformations. This increases effective dataset size and helps the model generalize to diverse retinal presentations.

2.2. Task 1: Transfer Learning Strategies

In the first task, two pretrained architectures, **EfficientNet** and **ResNet18**, were evaluated under three different transfer learning settings. In the no fine-tuning setup, pretrained weights were used directly for inference without any additional training. In the frozen backbone setting, the backbone weights were kept fixed while only the classifier was trained on the ODIR dataset. Finally, in the full fine-tuning configuration, both the backbone and classifier were jointly optimized. Performance was measured using the average F1-score computed across the three disease classes.

Table 1. Task 1 Offsite Test Performance (Average F-score)

Model	No FT	Frozen Backbone	Full FT
EfficientNet	0.63	0.69	0.80
ResNet18	0.61	0.63	0.82

Table 2. Task 1 Onsite Test Performance (Average F-score)

Model	Avg F-score
EfficientNet	0.81
ResNet18	0.82

2.3. Task 2: Loss Functions for Class Imbalance

To address the severe class imbalance, I explored two loss functions. **Focal Loss** (alpha=0.5, gamma=2.5) focuses on hard-to-classify samples, while **Class-Balanced Loss** (beta=0.9999) reweights the binary cross-entropy based on the training set class frequencies. The parameters were chosen empirically through small-scale validation experiments to maximize macro F1-score. As shown in Tables 3, Class-Balanced Loss consistently improved performance on underrepresented classes, particularly for ResNet18, which achieved the highest average F-score on the onsite test set.

2.4. Task 3: Attention Mechanisms

Two attention mechanisms were integrated into the selected backbones to improve feature representation:

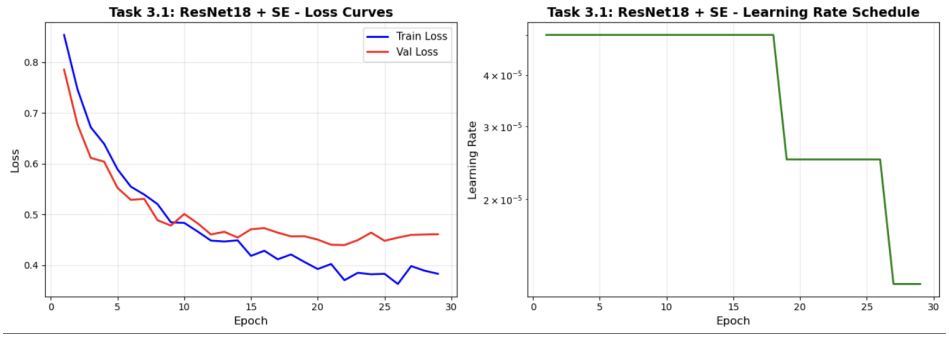

Fig. 1. Training and validation loss curves for ResNet18 with SE attention. The use of SE blocks together with dropout and L2 regularization helps stabilize training and reduces overfitting.

Table 3. Task 2 Offsite Test Performance (Average F-score)

Loss Function	DR	G	AMD	Avg F-score
ResNet18				
Focal Loss	0.78	0.84	0.77	0.80
Class-Balanced Loss	0.83	0.86	0.77	0.82
EfficientNet				
Focal Loss	0.76	0.78	0.77	0.77
Class-Balanced Loss	0.75	0.77	0.78	0.77

Table 4. ResNet 18 (Best Model) Task 2 Onsite Test F-score

Loss Function	Avg F-score
ResNet18 with Focal Loss	0.80
ResNet 18 with Class-Balanced Loss	0.83

- **Squeeze-and-Excitation (SE)** blocks, which adaptively recalibrate channel-wise feature responses.
- **Multi-head Attention (MHA)** captures long-range dependencies and interactions between feature channels.

Several combinations were evaluated, including SE, MHA, and SE+MHA heads.

Table 5. Task 3 Offsite Test Performance (Average F-score)

Model Variant	ResNet18	EfficientNet
+ SE	0.83	0.80
+ MHA	0.82	0.80
+ SE + MHA	0.83	0.80

Table 6. Task 3 Best Onsite Test Performance (F-score)

Model	Avg F-score
ResNet18 + SE	0.84
EfficientNet + SE + MHA	0.82

The results indicate that adding SE blocks provided the largest improvement on ResNet18, while combining SE and MHA offered marginal gains. EfficientNet variants benefited moderately from attention, with SE+MHA giving the best offsite F-score.

2.5. Task 4: Open Exploration and Performance Enhancement

In Task 4, I explored advanced architectures and learning strategies beyond the baseline CNN models to further improve multi-label retinal disease classification performance. The primary focus was on leveraging stronger backbone networks, improving generalization through regularization and data augmentation, and enhancing robustness using ensemble learning and model interpretability techniques.

*2.5.0.1. **Deeper Backbone Architectures***

I fine-tuned several high-capacity models pre-trained on ImageNet, including **Vision Transformer (ViT Base), ResNet50, DenseNet121, Inception V3,** and **ShuffleNet V2**. Among these, ViT Base achieved the strongest individual performance, highlighting the effectiveness of transformer-based global context modeling for retinal image analysis. CNN-based models such as ResNet50 and DenseNet121 also performed competitively, offering a favorable balance between performance and computational efficiency.

*2.5.0.2. **Advanced Training Strategies.***

To improve generalization and stability, I incorporated multiple training enhancements across all models. **Squeeze-and-Excitation (SE) blocks** were integrated to provide channel-wise attention, allowing the networks to emphasize informative feature maps. **MixUp augmentation** was applied to reduce overfitting and encourage smoother decision boundaries. To address class imbalance, **Focal Loss** with label smoothing was employed, which prioritizes hard samples while preventing overconfident predictions. Optimization was performed using **AdamW** with strong weight decay, combined with cosine annealing learning rate scheduling and warmup. A two-stage fine-tuning strategy was adopted, where the classifier

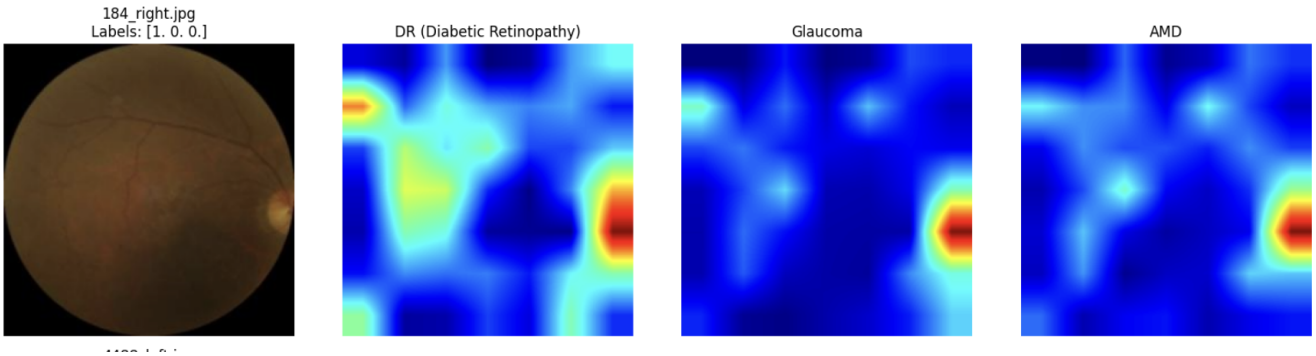

Fig. 2. GradCAM visualizations for retinal disease classification. Heatmaps highlight regions contributing most to model predictions (red indicates high importance). From left to right: Diabetic Retinopathy, Glaucoma, and AMD examples.

head was trained first, followed by full-network fine-tuning with a reduced learning rate.

*2.5.0.3. **Ensemble Learning***

To further boost robustness, I constructed an ensemble model using majority voting over the top-performing individual models: **ViT Base, ResNet50, and DenseNet121**. Ensemble learning leveraged the strengths of transformer and convolutional architectures, reducing variance and improving consistency across disease classes. The highest onsite F1 score was achieved by this ensemble i.e., **0.85**.

*2.5.0.4. **Explainable AI with GradCAM***

To enhance model interpretability, I implemented Gradient-weighted Class Activation Mapping (GradCAM) across CNN and transformer models. The generated **attention maps** confirmed that the models focus on clinically meaningful regions, such as the optic disc for glaucoma, and the macular region for AMD. These visual explanations validate that performance gains are driven by relevant anatomical features rather than spurious correlations.

3. RESULTS AND DISCUSSION

Experimental results demonstrate that stronger backbone architectures and advanced training strategies significantly improve retinal disease detection performance. Vision Transformers yielded the highest individual F1-Macro scores, while ensemble learning improved robustness. Attention mechanisms, regularization, and data augmentation played a critical role in mitigating overfitting, particularly for under-represented classes such as AMD. Overall, Task 4 highlights that combining architectural innovation, optimized training pipelines, and interpretability leads to meaningful and reliable performance improvements in multi-label medical image classification.

A key challenge throughout the experiments was the relatively small dataset size, which led to rapid overfitting. These models exhibited fast convergence followed by early performance saturation or degradation after only a few epochs. To mitigate this, extensive regularization techniques were employed, including strong weight decay, dropout, MixUp augmentation, and focal loss, enabling training over more epochs without severe overfitting.

Table 7. Onsite Test Performance (Average F1-score)

Model	**Avg F1-score**
Vision Transformer (ViT Base)	0.84
ResNet50	0.84
DenseNet121	0.83
Inception V3	0.79
ShuffleNet V2	0.80
Ensemble (Top 3 Models)	**0.85**

Table 7 summarizes the onsite test performance of all evaluated models. Among individual architectures, the Vision Transformer achieved the highest average F1-score, demonstrating the benefit of global context modeling for retinal disease classification.

4. CONCLUSION

This work presented a comprehensive study on multi-label retinal disease classification using deep learning. Through systematic evaluation of transfer learning strategies, loss functions, attention mechanisms, and advanced architectures, I demonstrated that careful model design and regularization are critical when working with limited medical imaging data. Vision Transformer models achieved the strongest individual performance, while ensemble learning improved robustness across disease classes. Despite persistent challenges related to overfitting and dataset size, the use of extensive augmentation, attention mechanisms, and explainable AI techniques resulted in clinically meaningful predictions.