TRANSFER LEARNING FOR MULTI-LABEL MEDICAL IMAGE CLASSIFICATION

Arshman Tariq, Marwa Bibi, Muhammad Faizan Tanveer

521153S-3006 Deep Learning, ITEE University of Oulu, Finland

1. INTRODUCTION

Retinal diseases, such as Diabetic Retinopathy (DR), Glaucoma (G), and Age-related Macular Degeneration (AMD), are causing vision impairment globally; however, obtaining large-scale, annotated medical datasets for training deep learning models remains a significant challenge. This project addresses the multi-label retinal disease classification problem using the ODIR dataset, which has a limited sample size and high class imbalance. The idea is to use transfer learning to overcome data scarcity and improve model performance.

2. METHODS

2.1. Dataset and problem setting

The ODIR dataset focuses on: Diabetic Retinopathy (DR), Glaucoma (G), and Age-related Macular Degeneration (AMD). Each image can exhibit none, one, or multiple conditions. It contains highly imbalanced disease prevalence, see Figure 1, which bias standard training toward majority labels.

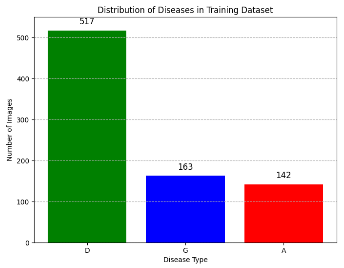


Fig. 1. Imbalanced Class distribution of training set.

2.2. Model architectures for transfer learning

ResNet18 and EfficientNet-B0 backbones are loaded from pre-trained checkpoints and adapted for 3-label prediction.

2.3. Training strategies and hyperparameters

Training is performed in the following stages:

- **Task 1.2 (frozen backbone)**: The backbone is frozen, and only the classifier head is trained using Adam.

- **Task 1.3 (full fine-tuning)**: All layers are unfrozen and trained using Adam with learning rate 2×10^{-5} to avoid destroying pretrained representations.

- **Task 2 (loss functions for imbalance)**: Full fine-tuning with Adam and learning rate 2×10^{-5}, while replacing BCE with (i) focal loss and (ii) class-balanced BCE to mitigate imbalance effects.

- **Task 3 (attention models)**: Attention-augmented ResNet/EfficientNet models (SE or MHA) are trained with the Task 3 loss and Adam at learning rate 2×10^{-5}.

2.4. Loss functions

2.4.0.1. Binary cross-entropy (BCE) with logits:
A baseline multi-label loss functions, applied per class.

2.4.0.2. Focal loss (multi-label):
Focal loss down-weights easy negatives and focuses learning on hard samples, implemented as a multi-label focal loss with focusing parameter γ and balancing factor α.

2.4.0.3. Class-balanced BCE:
Class-balanced BCE weights positive labels using the effective-number-of-samples approach, yielding per-class positive weights that reduce majority-class dominance.

2.5. Attention mechanisms (SE and MHA)

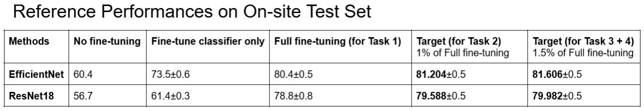
Attention was introduced to focus on clinically relevant retinal structures under limited and imbalanced data.

1. **Squeeze-and-Excitation (SE)** performs channel-wise reweighting by learning per-channel gates from global pooled features, improving feature selection with minimal overhead.

2. **Multi-Head Attention (MHA)** models long-range dependencies by computing attention weights across spatial features, enabling the model to relate distant regions when forming predictions.

2.6. Evaluation protocol

Performance is reported using per-class F-score and checkpointed using the best validation macro F-score, which is then evaluated on the off-site split and used to generate the on-site submission CSV. The reference onsite scores are in Figure 2.



Reference Performances on On-site Test Set

Methods	No fine-tuning	Fine-tune classifier only	Full fine-tuning (for Task 1)	Target (for Task 2) 1% of Full fine-tuning	Target (for Task 3 + 4) 1.5% of Full fine-tuning
EfficientNet	60.4	73.5±0.6	80.4±0.5	**81.204**±0.5	**81.606**±0.5
ResNet18	56.7	61.4±0.3	78.8±0.8	**79.588**±0.5	**79.982**±0.5

Fig. 2. Reference performances on the onsite test set.

2.7. VAE-based data augmentation - (Task 4a)

For Open Question task, three modules were implemented to have an in-depth analysis and performance comaprison, the first one was focused on mitigating data scarcity. A Variational Autoencoder (VAE) was trained to learn the training distribution and generate additional synthetic images for augmentation while keeping the training pipeline unchanged. The data distribution and augmented images can be seen here.

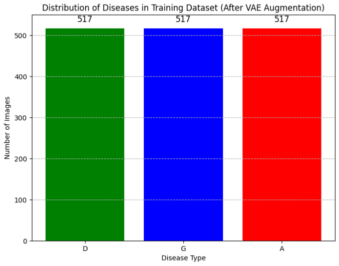
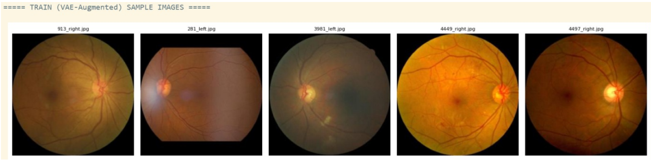


Fig. 3. VAE balanced Class distribution for DR, Glaucoma, and AMD in the training set.



```
===== TRAIN (VAE-Augmented) SAMPLE IMAGES =====
```

Fig. 4. Augmented Train Dataset.

2.8. Stronger backbones (ViT and Swin) - (Task 4b)

Beyond CNN backbones, transformer-based encoders for richer global representation learning were also evaluated.

1. **Vision Transformer (ViT)** splits an image into patches (tokens) and applies self-attention over tokens, which can capture global context but typically benefits from more data or strong pretraining.
2. **Swin Transformer** introduces hierarchical, window-based self-attention with shifted windows, improving scalability and providing multi-scale features that are often effective for medical imaging.

2.9. GradCAM visualisation and guided training

2.9.0.1. GradCAM Visualisation.

GradCAM generates a class-specific heatmap by backpropagating gradients to the last convolution layer and combining them with the corresponding activations to highlight important features in the images. This makes the black box frameworks interpretable which is essential for medical applications, as they build confidence and support clinical decision.

2.9.0.2. GradCAM-guided training. - (Task 4c)

CAM-based regularizer was added to the classification loss to shape where the model attends, using an entropy-style term. This improved interpretability, but reduced accuracy as it encourages overly diffuse attention.

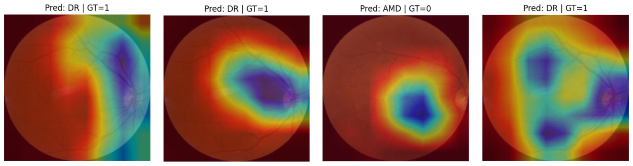


Fig. 5. Heatmap using GradCAM for important features.

3. RESULTS

3.1. Baseline transfer learning (Task 1)

Full fine-tuning (Task 1.3) significantly improves macro F-score over frozen-backbone training, confirming the value of adapting deep features to retinal domain statistics despite limited data. For example, ResNet Task 1.2 reaches an offsite macro F-score of 0.5313, while ResNet Task 1.3 reaches average F score of 0.7726.

Table 1. Off-site (ResNet18) scores for Task 1.2 and 1.3.

Offsite Testing	Class	F-score	Precision	Recall
ResNet Task 1.2	DR	0.7626	0.6622	0.8991
	Glaucoma	0.5349	0.8214	0.3966
	AMD	0.2963	0.8000	0.1818
ResNet Task 1.3	DR	0.8356	0.8103	0.8624
	Glaucoma	0.8073	0.8627	0.7586
	AMD	0.6750	0.7500	0.6136

3.2. Loss functions for imbalance (Task 2)

Loss re-design changes per-class trade-offs, particularly improving minority-class learning when compared to plain BCE training, though behavior varies by backbone.

3.3. Attention mechanisms (Task 3)

Adding attention improves performance by encouraging more informative feature weighting. For example, ResNet Task 2.1 reaches an offsite macro F-score of 0.7811, while ResNet Task 1.3 reaches average F score of 0.7786.

Table 2. Off-site (ResNet18) scores for Task 2.1 and 3.2.

Offsite Testing	Class	F-score	Precision	Recall
Task 2.1 (Focal)	DR	0.9058	0.9191	0.8929
	Glaucoma	0.7556	0.8293	0.6939
	AMD	0.6818	0.6818	0.6818
Task 3.2 (MHA)	DR	0.8208	0.8447	0.7982
	Glaucoma	0.8269	0.9348	0.7414
	AMD	0.6882	0.6531	0.7273

3.4. VAE, Swin & Vit and GradCAM (Task 4)

Table 3 compares off-site macro F-score for attention (MHA on ResNet), data augmentation (VAE + MHA), stronger backbones (ViT and Swin), and GradCAM-guided training. MHA improves performance by modeling long-range dependencies, while VAE augmentation can help data scarcity but may introduce distribution shift that reduces macro F-score. Transformer backbones (ViT/Swin) achieve the strongest macro F-scores, suggesting that global self-attention is beneficial for multi-label retinal classification, whereas GradCAM-guided training trades accuracy for more interpretable attention maps, leading to a lower macro F-score.

Table 3. Off-site average F-scores for Task 4 variants.

Method	Avg F-score
Task 3.2 MHA (ResNet)	0.7786
Task 4 VAE + MHA (ResNet)	0.7587
Task 4 ViT	0.7968
Task 4 Swin	0.7749
Task 4 GradCAM + MHA (ResNet)	0.7138

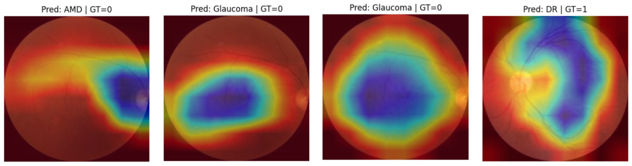


Fig. 6. Heatmaps after GradCAM guided training.

3.5. On-site Performance Summary

Table 4 summarizes the on-site test performance trend across the explored training strategies (fine-tuning, imbalance-aware losses, attention, and open-question variants), as provided in the on-site score sheet.

Table 4. On-site test performance summary (ResNet18).

Method / Task	On-site score
No fine-tuning	56.668
Fine-tune classifier only (Task 1.2)	61.089
Full fine-tuning (Task 1.3)	83.14
Reference full fine-tuning	**78.8** ± 0.8
Target (1% reference fine-tuning)	**79.588** ± 0.5
Loss function variants (Task 2)	82.168 / 83.216
Target (1.5% reference fine-tuning)	**79.982** ± 0.5
Attention mechanisms (Task 3)	81.554 / 82.686
VAE augmentation (Task 4a)	80.984 / 82.40
ViT & Swin (Task 4b)	85.717 / 85.122
GradCAM-guided training (Task 4c)	75.992 / 73.788

4. DISCUSSION

4.1. Key takeaways

Full fine-tuning worked best, because the backbone must adapt beyond the classifier head for retinal images. While later tasks mainly refined generalization.

4.2. Augmentation vs. transformers

Synthetic images using VAE augmentation may not preserve fine diagnostic structure. Meanwhile, ViT/Swin worked best because global self-attention supports multi-label decisions.

4.3. Why GradCAM-guided training reduced accuracy

GradCAM regularizer explicitly pushes saliency maps toward broader, smoother activation, which can weaken sharp discriminative signals needed for classification.

5. TEAM CONTRIBUTION AND GITHUB LINK

Table 5. Team members contribution by task.

Task / Component	Contributor(s)
Task 1 (Fine-tuning)	Marwa Bibi
Task 2 (Loss Function)	Arshman Tariq
Task 3 (Attention Mechanisms)	Muhammad Faizan Tanveer
Task 4 (VAE Augmentation)	Marwa Bibi
Task 4 (ViT & Swin Transformer)	Arshman Tariq; Muhammad Faizan Tanveer
Task 4 (GradCAM)	Arshman Tariq; Muhammad Faizan Tanveer
Report Writing	Arshman Tariq; Marwa Bibi; Muhammad Faizan Tanveer

https://github.com/FaizanTanwir/521153S-3006-DL-Transfer-Learning-for-Multi-label-Medical-Image-Classification.git