Transfer Learning for Multi-label Retinal Disease Classification

Kaggle group name: DLsns

Shokooh Mirfakhraei
smirfakh25@student.oulu.fi
Student ID: 2510955

Negin Hadad
nhadad25@student.oulu.fi
Student ID: 2509733

Seyedehsahar Fatemi Abhari
sfatemia25@student.oulu.fi
Student ID: 2508265

1. Introduction

Automated diagnosis of retinal diseases is critical given the global shortage of ophthalmologists. However, medical imaging datasets often suffer from limited size and severe class imbalance. This project addresses the multi-label classification of DR, Glaucoma, and AMD by systematically optimizing a deep learning pipeline.

The primary objectives are:

1. **Transfer Learning Optimization:** To determine the optimal depth of fine-tuning for adapting ImageNet/medical-pretrained weights to the ODIR domain.
2. **Imbalance Mitigation:** To evaluate loss functions that penalize easy negatives and re-weight minority classes.
3. **Architectural Enhancement:** To integrate channel-wise attention mechanisms for improved feature discrimination.
4. **Ensemble Learning:** To assess whether combining diverse models from Tasks 1-3 improves generalization beyond single-model performance.

2. Methodology

2.1 Dataset and Preprocessing

In addition to all the preprocessing performed on the ODIR dataset, we utilised augmentation to mitigate overfitting on the small training set (N = 800). We applied random horizontal/vertical flips (p=0.5), random rotations (15 degrees), and colour jitter (brightness/contrast adaptation).

2.2 Experimental Protocols

Task 1: Transfer Learning Paradigms

Three initialisation strategies were evaluated to assess domain adaptation:

- **Protocol A (No Fine-tuning):** Backbone weights frozen; only the final linear classifier was trained.
- **Protocol B (Frozen Backbone):** Backbone frozen; classifier trained with high learning rate (1e-3).
- **Protocol C (Full Fine-tuning):** All layers unfrozen; trained with reduced learning rate (1e-4) to preserve pretrained features while adapting to ODIR statistics. In this task we evaluate two convolutional neural network backbones: ResNet18 and EfficientNet[1][2]

All models were trained for 50 epochs with batch size 32.

Task 2: Loss Function Analysis

The training set exhibits severe class imbalance (DR: 537, Glaucoma: 163, AMD: 142 samples). Standard Binary Cross-Entropy Loss tends to be dominated by the majority class. We implemented two specialized loss functions:

- **Focal Loss (α=0.75, γ=2.5):** Down-weights well-classified examples via modulating factor
- $\left(1 - p_t\right)^{\gamma}$, forcing the model to focus on hard negatives and minority classes [3].
- **Class-Balanced Loss (β=0.9999):** Re-weights each class by effective sample number: $w_i = (1 - \beta) / (1 - \beta^{n_i})$, where n_i is the sample count for class i. This compensates for frequency imbalance[4].

Both models used ResNet18 initialized from Task 1.3 weights, trained with AdamW optimizer (lr=2e-5, weight decay=1e-3), OneCycleLR scheduler, mixed precision training, and aggressive augmentation including Gaussian blur (σ=0.1-2.0), random erasing (p=0.4), and affine transformations (60 epochs, batch size 12).

Task 3: Attention Mechanisms

In this task, we evaluated whether adding attention mechanisms improves multi-label retinal disease classification. We used ResNet18 with full fine-tuning from Task 1.3 as the baseline model. Attention-based models were initialized using the pretrained weights from Task 1.3 to ensure a fair comparison.

3.1. Squeeze-and-Excitation (SE) Block

The Squeeze-and-Excitation block was added after the final convolutional layer 4 of ResNet18. The SE module recalibrates channel-wise features by first applying global average pooling and then learning channel importance using two fully connected layers with a sigmoid activation[5]. The SE-based model was trained for 50 epochs with a batch size of 32 and a learning rate of 1e - 5. All images were resized to 256. During inference, test-time augmentation (TTA) was applied by averaging predictions across multiple augmented versions of each image to improve robustness.

3.2. Multi-Head Attention (MHA)

We also tested Multi-Head Self-Attention (MHA) with 8 attention heads. This attention module was designed to capture global relationships across the image[6]. To reduce computation, the feature map size was reduced before applying attention. Layer normalization and residual connections were used to stabilize training.

Task 4: Ensemble Learning

Rather than training new models, we leveraged all previously trained checkpoints from Tasks 1-3 to create a diverse ensemble. The ensemble pool consisted of 12 base models: Task 1 (ResNet18 + EfficientNet with 3 fine-tuning protocols = 6 models), Task 2 (ResNet18 + EfficientNet with Focal Loss and Class-Balanced Loss = 4 models), and Task 3 (ResNet18 with SE and MHA attention = 2 models).

Each checkpoint required task-specific loading: Task 1 models used standard classification heads; Task 2 models included dropout layers (p=0.5) before the final linear layer; Task 3 models incorporated attention modules (SE with reduction=16 or MHA with 8 heads) after ResNet18's layer4. All models output three logits converted to probabilities via sigmoid activation.

Four ensemble strategies were evaluated:

- **Weighted Average Ensemble (Soft Voting):** Equal-weight averaging (1/12 per model) of all model probabilities with fixed threshold (0.5). Two variants were tested: (1) standard inference and (2) test-time augmentation (TTA) averaging predictions across original, horizontal-flip, and vertical-flip transformations.
- **Stacking with Random Forest:** Each base model's 3-class probabilities were concatenated into 36-dimensional meta-features (12 models × 3 classes). Separate Random Forest classifiers (100 trees, max depth=10) were trained per disease on training set meta-features.
- **Adaptive Threshold Optimization:** Weighted ensemble probabilities were generated, then per-class decision thresholds were optimized on the validation set via grid search (range: 0.25-0.75, step: 0.05) to maximize individual F1-scores for DR, Glaucoma, and AMD independently.

All methods were evaluated on the offsite test set using average F1-score across the three diseases.

3. Experiments and Results

3.1 Task 1: Transfer Learning (Onsite Test Set)

We first evaluated the generalisation capability of the models on the unseen Onsite Test Set (Kaggle).

Table 1: Transfer Learning Results (Onsite Test Set)

Backbone	**Protocol**	**Target F1-Score**	**Our F1-Score**
ResNet18	No Fine-tuning	56.7	56.7[1]
	Frozen Backbone	61.4 ± 0.3	64.5
	Full Fine-tuning	78.8 ± 0.8	81.8
EfficientNet	No Fine-tuning	60.4	60.4
	Frozen Backbone	73.5% ± 0.6	74.6
	Full Fine-tuning	80.4% ± 0.5	80.4

Observation: Protocol C (Full Fine-tuning) achieved the highest performance. The significant gap between Protocol B (64.5.0%) and C (81.8%) indicates that the pretrained features, while robust, required adaptation to the specific visual characteristics of the ODIR dataset.

3.2 Task 1: Detailed Offsite Performance

Based on the comparative analysis, **ResNet18 with Full Fine-Tuning** was identified as our optimal model, achieving the highest Average F1-score of 83.19% on the Offsite Test Set, slightly outperforming EfficientNet (81.60%). **Table 2** details the performance metrics per disease for this best-performing configuration.

Table 2: Detailed Offsite Parameters (ResNet18)

Protocol	**Disease**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Kappa**
Full Fine-Tuning	Diabetic Retinopathy	84.5	81.34	84.17	82.41	64.93
	Glaucoma	89.5	87.24	83.4	85.06	70.18
	AMD	93	82.12	82.12	82.12	64.25
	Average	**89**	**83.56**	**83.23**	**83.19**	**66.45**
Frozen Backbone	Diabetic Retinopathy	81.5	79.06	74.88	76.4	53.05
	Glaucoma	82	81.54	66.02	68.83	39.71
	AMD	91.05	82.66	67.34	71.93	44.41
	Average	**84.85**	**81.08**	**69.41**	**72.38**	**45.72**
No Fine-Tuning	Diabetic Retinopathy	51.5	51.7	52.02	49.58	3.39
	Glaucoma	78.5	70.63	67.84	68.93	38.04
	AMD	78.5	63.05	75.97	64.72	32.11
	Average	**69.5**	**61.79**	**65.27**	**61.07**	**24.51**

Table 3: Detailed Offsite Parameters (EfficientNet)

Protocol	**Disease**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Kappa**
Full Fine-Tuning	Diabetic Retinopathy	83.5	80.4	83.93	81.55	63.33
	Glaucoma	87	82.01	84.5	83.11	66.26
	AMD	91	76.86	84.98	80.11	60.35
	Average	**87.16**	**79.75**	**84.47**	**81.6**	**63.31**
Frozen Backbone	Diabetic Retinopathy	78.5	74.41	74.64	74.53	49.05
	Glaucoma	85	81.4	75.59	77.78	55.8
	AMD	94	90.17	76.71	81.68	63.55
	Average	**85.83**	**82**	**75.64**	**78**	**56.13**
No Fine-Tuning	Diabetic Retinopathy	60	55.88	56.67	55.75	12.28
	Glaucoma	79.5	72.43	73.33	82.85	45.71
	AMD	71.5	60.41	74.03	59.46	24.82
	Average	**70.33**	**62.9**	**68.01**	**66.02**	**27.6**

Performance Analysis

- **Superiority of Full Fine-Tuning:** The results conclusively demonstrate that full fine-tuning is essential for this task. The Average F1-score improved dramatically from 61.07% (No Fine-Tuning) and 72.38% (Frozen Backbone) to 83.19% (Full Fine-Tuning). This confirms that the deep features learned from ImageNet/Medical pre-training required significant adaptation to capture the specific nuances of the ODIR dataset.
- **Robust Minority Class Detection:** Contrary to typical expectations in imbalanced datasets, our model achieved its highest F1-score on Glaucoma (85.06%), despite it being a minority class compared to Diabetic Retinopathy. This suggests that the distinct visual features of Glaucoma (such as optic disc cupping) were effectively learned by the ResNet18 backbone during the fine-tuning process.
- **Balanced Performance Across Diseases:** The model exhibits remarkable consistency, with F1-scores for all three diseases clustering closely between 82.1% and 85.1%. This indicates that the Full Fine-Tuning strategy successfully mitigated the risk of the model biasing heavily toward the majority class (DR), resulting in a well-generalized solution for all three conditions.

3.3 Impact of Loss Functions (Task 2)

We evaluated both loss functions using ResNet18 and EfficientNet backbones. Table 4 presents the offsite test performance.

Table 4: Loss Function Comparison (Offsite Test Set)

Backbone	Protocol	Disease	Precision	Recall	F1-score
ResNet18	Focal Loss	Diabetic Retinopathy	85.81	95.00	90.17
		Glaucoma	82.22	75.51	78.72
		AMD	42.86	68.18	52.63
		Average	70.29	79.56	73.84
ResNet18	Class-Balanced Loss	Diabetic Retinopathy	93.80	86.43	89.96
		Glaucoma	84.78	79.59	82.11
		AMD	66.67	63.64	65.12
		Average	81.75	76.55	79.06
EfficientNet	Focal Loss	Diabetic Retinopathy	84.77	91.43	87.97
		Glaucoma	68.97	81.63	74.77
		AMD	53.33	72.73	61.54
		Average	69.02	81.93	74.76
EfficientNet	Class-Balanced Loss	Diabetic Retinopathy	87.50	85.00	86.23
		Glaucoma	78.26	73.47	75.79
		AMD	51.85	63.64	57.14
		Average	72.54	74.04	73.05

Performance Analysis

- **Effectiveness of Class-Balanced Loss:** The results clearly demonstrate that Class-Balanced Loss is more effective than Focal Loss for addressing class imbalance in multi-label retinal disease classification. Using the ResNet18 backbone, the Offsite Average F1-score improved from 73.84% with Focal Loss to 79.06% with Class-Balanced Loss. This improvement indicates that re-weighting the loss according to class frequency enables the model to better capture underrepresented disease patterns while preserving strong overall performance. The consistent increase in the average F1-score confirms the advantage of Class-Balanced Loss in achieving a more balanced and reliable classification across all disease categories.
- **Improved Minority Class Recognition:**

A key benefit of Class-Balanced Loss is its strong impact on minority classes. While the F1-score for Diabetic Retinopathy (the majority class) remained high and stable (89.96%), the performance on minority diseases improved significantly. For example, the Glaucoma F1-score increased from 78.72% to 82.11%. This confirms that Class-Balanced Loss successfully mitigates bias toward the majority class and enables more effective learning of underrepresented diseases.

- **Generalization to Onsite Test Set:**

The superiority of Class-Balanced Loss is further validated by the Kaggle onsite test results. The ResNet18 model trained with Class-Balanced Loss achieved an Onsite Average F1-score of 82.694%, which outperforms both the Focal Loss configuration (81.440%) and all EfficientNet-based models. Compared to the reference full fine-tuning onsite performance reported in the project instructions (approximately 80.4%), this result represents an improvement of more than +1.0%.

Based on offsite and onsite results, ResNet18 with Class-Balanced Loss was selected as the best model for Task 2. It achieved strong overall performance, improved minority class recognition, and exceeded the reference onsite score. Table 5 represents the onsite result for each Loss function.

Table 5: Onsite Test Set Performance (Task 2 – Loss Functions)

Backbone	Loss Function	Onsite Avg F1 (%)
ResNet18	Focal Loss	81.44
ResNet18	Class-Balanced Loss	82.69
EfficientNet	Focal Loss	80.18
EfficientNet	Class-Balanced Loss	82.38

3. Attention Mechanism Evaluation (Task 3)

Both attention mechanisms were integrated into ResNet18 using Task 1.3 weights as initialization. Results are shown in Table 6.

Table 6: Attention Mechanism Performance- Offsite

Protocol	Disease	Accuracy	Precision	Recall	F1-score

[1] *For consistency and ease of interpretation, all performance metrics are expressed as percentages.*

SE Block	Diabetic Retinopathy	84.00	90.30	86.43	88.32
	Glaucoma	91.50	88.10	75.51	81.32
	AMD	93.50	73.68	63.64	68.29
	Average	89.67	84.03	75.19	79.31
MHA	Diabetic Retinopathy	85.00	89.86	88.57	89.21
	Glaucoma	88.50	75.00	79.59	77.23
	AMD	93.00	68.18	68.18	68.18
	Average	88.83	77.68	78.78	78.21

Analysis: The SE Block achieved higher overall performance than MHA, with an average F1-score of 79.31%. SE showed strong precision and recall for Diabetic Retinopathy, resulting in the highest DR F1-score among the attention models. It also achieved better balance between precision and recall for Glaucoma compared to MHA. For AMD, both models showed lower recall, which limited the overall F1-score.

The MHA model achieved an average F1-score of 78.21%. While MHA reached slightly higher recall for Diabetic Retinopathy, it showed lower precision and recall for Glaucoma compared to SE. Performance on AMD was similar for both models, indicating that neither attention mechanism effectively improved AMD detection on the offsite test set.

In terms of accuracy, both attention-based models achieved high values across all diseases, especially for AMD. However, the lower recall for minority classes reduced the final F1-scores, explaining why the overall performance did not surpass the full fine-tuning baseline.

The onsite test performance is summarized in Table 7. The full fine-tuning baseline from Task 1 (ResNet18, Task 1.3) achieved an onsite average F1-score of 81.8%.

Both attention-based models improved upon this baseline. The SE Block achieved the highest onsite performance with an average F1-score of 83.703%, while MHA achieved 82.807%. This shows that attention mechanisms improved generalization on the unseen onsite test set, even though offsite performance did not exceed the baseline.

Table 7: Onsite Test Set Performance

Model	Onsite Avg F1
Full Fine-Tuning (Task 1.3)	81.8
SE Block	83.703
MHA	82.807

3.4 Ensemble Learning Results (Task 4)

Several ensemble strategies were evaluated by combining models from Tasks 1–3, including weighted averaging, stacking with a Random Forest, and adaptive thresholding. On the offsite test set, the Ultimate Weighted Ensemble without test-time augmentation (TTA) achieved the best performance (80.66% average F1).

On the onsite test set, ensemble methods further improved generalization. The Ultimate Weighted Ensemble without TTA achieved the highest onsite F1-score of 83.66%, outperforming both stacking-based and adaptive threshold ensembles. Based on these results, this method was selected as the final ensemble model.

Table 8: Performance Comparison of Ultimate Ensemble Methods - offsite

Method	Test-Time Augmentation (TTA)	DR F1	Glaucoma F1	AMD F1	Offsite Avg F1
Ultimate Weighted Ensemble	Yes	91.49	74.73	75.00	80.40
Ultimate Weighted Ensemble	No	91.43	75.56	75.00	80.66
Ultimate Stacking Ensemble (Random Forest)	No	90.77	79.17	71.11	80.35
Ultimate Adaptive Threshold	No	90.51	78.43	71.43	80.12

Table 9: Performance of Ultimate Ensemble Methods- Onsite

Method	Offsite Avg F1 (%)	Onsite Avg F1 (%)
Ultimate Weighted Ensemble (No TTA)	80.66	83.66
Ultimate Weighted Ensemble (With TTA)	80.40	83.30
Ultimate Stacking Ensemble	80.35	83.28
Ultimate Adaptive Threshold	80.12	83.19

4. Discussion and Conclusion

This study systematically optimized a deep learning pipeline for multi-label retinal disease classification through transfer learning, loss function design, attention mechanisms, and ensemble learning.

1. **Transfer Learning Effectiveness:**
   The results demonstrate that ResNet18 with full fine-tuning is the most effective backbone for this task, outperforming EfficientNet by approximately 1.6% in average F1-score on the offsite test set (83.19% vs. 81.60%). Although EfficientNet showed stronger performance under the no fine-tuning setting, ResNet18 benefited more from full network adaptation, indicating better domain transferability.
2. **Handling Class Imbalance:**
   Class imbalance was addressed using Focal Loss and Class-Balanced Loss. The experiments showed that Class-Balanced Loss consistently outperformed Focal Loss, particularly for minority classes such as Glaucoma and AMD. As a result, ResNet18 achieved balanced performance across diseases, with Glaucoma and AMD matching or exceeding the performance of the majority class, Diabetic Retinopathy. This confirms that re-weighting samples based on effective class frequency is well suited for this task.
3. **Impact of Attention Mechanisms:**
   Channel-wise attention using Squeeze-and-Excitation (SE) improved disease-specific detection and onsite generalization but did not surpass the baseline on the offsite test set. In contrast, Multi-Head Attention (MHA) introduced higher complexity and showed weaker generalization, particularly for AMD. These findings suggest that attention mechanisms require careful design when training data is limited.
4. **Ensemble Learning and Generalization:**
   Ensemble methods combining models from Tasks 1–3 further improved robustness and generalization. The ultimate weighted ensemble without test-time augmentation achieved the best offsite performance, while ensemble models consistently improved onsite results. This highlights the effectiveness of ensemble learning for improving final performance on unseen data.

5. Statement of Contributions

Negin Hadad: Led the implementation of Task 1 (Transfer Learning), including the training and evaluation of the ResNet18 and EfficientNet backbones. Conducted the ablation study comparing 'No Fine-tuning,' 'Frozen Backbone,' and 'Full Fine-tuning' strategies. Additionally, implemented the Ensemble Learning method (Task 4), and established the project code structure by integrating Tasks 1 through 4. Moreover, set up the report template, authored the technical report for Task 1, and finally, created the project README.md file.

Seyedehsahar Fatemi Abhari: Led the implementation of Task 2 (Loss Function Design for Class Imbalance). Developed and trained both ResNet18 and EfficientNet models using Focal Loss and Class-Balanced Loss, including custom loss implementations, advanced data augmentation strategies, full fine-tuning, early stopping, and comprehensive offsite and onsite evaluations. Conducted a comparative analysis of loss functions. Authored the Task 2 section of the technical report. Additionally, implemented Task 4 (Ensemble Learning) by designing and evaluating multiple ensemble strategies, including weighted averaging, max voting, stacking with meta-learners, adaptive threshold optimization, and advanced multi-task ensembles integrating models from Tasks 1, 2, and 3.

Shokooh Mirfakhraei: Was responsible for Task 3 (Attention Mechanisms), including the implementation and evaluation of attention-augmented ResNet18 models using Squeeze-and-Excitation (SE) and Multi-Head Attention (MHA). Defined the training and evaluation methodology for Task 3 and conducted comprehensive offsite and onsite comparisons against the full fine-tuning baseline. Authored the Methodology of task 3 and Results and discussion sections for Task 3 and Task 4, and handled the overall report formatting, ensuring compliance with IEEE/ICIP guidelines.

6. References:

[1] K. He *et al*., "Deep Residual Learning for Image Recognition," *Proc. CVPR*, 2016.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for CNNs," *Proc. ICML*, 2019.

[3] T.-Y. Lin *et al*., "Focal Loss for Dense Object Detection," *Proc. ICCV*, 2017.

[4] Y. Cui *et al*., "Class-Balanced Loss Based on Effective Number of Samples," *Proc. CVPR*, 2019.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Proc. CVPR*, 2018.

[6] A. Vaswani *et al*., "Attention Is All You Need," *NeurIPS*, 2017.