Beyond Baselines: Improving Multi-Label Retinal Disease Detection Over EfficientNet (60.4 F1) and ResNet18 (56.7 F1)

Brian Kiprop Kibor
University of Oulu
brian.kibor@student.oulu.fi

Md Saddam Hossain Remus
University of Oulu

Shamim Ahamed
University of Oulu
shamim.ahamed@student.oulu.fi

1. ABSTRACT

Multi-label retinal disease detection is a major concern particularly for medical diagnosis fraternity. Conditions such as Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD) and Glaucoma are the main subject areas of discussion as they pose the significant risk of irreversible vision loss. Even though heavy research has been going on in this field, the output has been shortcoming due to limited size and huge annotation cost associated with medical imaging datasets. This project addresses these challenges by leveraging transfer learning with a pretrained convolutional neural network (CNNs) and fine tuning the models for multi-label classification on the ODIR dataset.

Through investigation of different design strategies affecting performance of deep learning models i.e. transfer learning strategies, loss function for addressing label imbalance and attention mechanism we demonstrate the changes in performance and interpretability of the models. Additionally, we also analyze preprocessing techniques such as ensemble methods, GradCAM to improve key metrics.

Our work demonstrates that careful fine tuning, appropriate loss function and attention integration to deep learning models improve performance significantly surpassing the baseline models. Overall, the study provided us with deep insights on building reliable deep learning pipelines that can be extended to any field.

Index Terms— Multi-label classification, Retinal disease detection, Transfer learning, Attention mechanisms.

2. INTRODUCTION

According to the World Health Organization, at least 2.2 billion people worldwide live with vision impairment or blindness, with retinal diseases such as DR, AMD, and Glaucoma recognized as major contributors to irreversible vision loss [1]]. Diabetic Retinopathy (DR) a diabetes complication often stems from prolonged blood glucose level which when elevated damage the small blood vessels in the retina. This leads to vision impairment and blindness if diagnosis and treatment is not done early enough. Age-related Macular Degeneration (AMD) attacks the macula portion of the eye primarily impairing central vision [2]. On the other hand, Glaucoma causes optical nerve damage leading to elevated intraocular pressure. With time this leads to gradual and irreversible peripheral vision loss.

As in all medical fields, early diagnosis is essential to mitigating and treating these conditions, as timely intervention can slow the disease progression. However, manual examination of retinal images is time consuming, costly and require expert knowledge which is limited globally (1.7 Doctors per 1000 people) [3]. Additionally, manual assessment is prone to inter-observer variability.

Recent advances in Machine Learning field has enabled the automation of image classification offering great tools for the detection and classification of retinal conditions. However, developing reliable medical imaging models has remained a challenge because of the limited labeled data, high annotation cost, privacy concerns, class imbalance and non-standardized data. This project addresses these challenges through the use of transfer learning on a pretrained CNN model together with image augmentation techniques to improve key metrics for multi-label retinal disease detection.

3. APPROACH

3.1. Data Preparation

This project utilizes the publicly available ODIR retinal image dataset from Kaggle to perform effective transfer learning. The dataset contains fundus images of Diabetic Retinopathy (DR), Glaucoma (G), and Age-related Macular Degeneration (AMD). The dataset is split into 800 training, 200 validation, 300 offsite and 250 onsite test images.

The baseline models (EfficientNet and ResNet18) utilized have been trained on ADAM[4], REFUGE2[5], and APTOS[6] datasets. These datasets cover wide range of multi-label retinal diseases varying in severity providing an accurate baseline model to be expanded.

All images used for this work are resized to fixed resolution 256×256 and normalized using ImageNet statistics to match the model input requirements. Dataset imbalance is addressed at the loss-function level as opposed to data resampling. We employed Focal Loss and Class Balanced Loss.

3.2. Evaluation Metrics

Given the inherent class imbalance typical of medical imaging datasets and the clinical importance of minimizing diagnostic errors, accuracy alone is insufficient in choosing an appropriate model [7]. Our analysis will focus primarily on F1-score complemented by precision, recall and Cohen's Kappa Score to assess the model performance.

i. Precision

This is the proportion of correct positives amongst all predicted positives. Indicates model's ability to avoid false positives hence avoiding false alarms in retinal disease diagnostics which can lead to unnecessary and expensive follow ups.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

ii. Recall

Model sensitivity. Quantifies proportion of actual positive diagnostics identified by the model correctly. A low recall is dangerous as it implies that the model is has a lot of false negatives as serious illness can remain undiagnosed.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

iii. F1-score

Harmonic mean of precision and recall, balancing both false positives and false negatives.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Useful in imbalanced datasets as it helps avoid bias caused by the imbalance.

iv. Cohen's Kappa

Quantifies the agreement between the predicted labels and ground truths while accounting for agreement occurring by chance.

Kappa is provided by:

$$\text{K} = \frac{po - pe}{1 - pe} \quad (4)$$

Where po = Observed Agreement pe = Expected Agreement. Values closer to 1 are preferred as they indicate stronger agreement beyond chance. Hence useful in accessing reliability and robustness of models.

3.3. Model Training

Transfer learning is widely utilized in medical image analysis to address the unavailability of labelled data. Models pre-trained on large scale datasets such as ImageNet are leveraged and generalized to new domains and fine-tuned as fit. Below we discuss the different strategies.

3.3.1. No Fine Tuning

The pre-trained CNN backbone is used solely as a fixed feature extractor. Backbone parameters are fixed and no weight updates. Assumes that features learned from ImageNet samples are sufficient to classify new samples. Computationally efficient and less, however lower performance due to missing domain specific features [8].

3.3.2. Fine Tuning Classifier Only

Here, the CNN backbone remains frozen while the classification layers are trained on the domain specific samples. As a result, decision boundaries to new labels are adjusted while the pre-trained features are maintained. Effective when dataset size is moderate and low-level features remain transferable across domains [9]. Nonetheless, its performance even though exceeds no fine tuning, it still misses specific features.

3.3.3. Full-fine Tuning

Both the backbone and classifier weights are updated during training using the new domain samples. Early CNN layers adjust to new domain specific features while deeper layers learn discriminative features giving it improved performance [9]. However, increased computation and overfitting risk more so in class imbalanced datasets.

3.3.4. Loss Mechanism

As mentioned, medical imaging datasets are inherently class imbalanced and have overlapping patterns hence necessity of loss functions; Binary Cross-Entropy (BCE), treats each label as an independent binary classification task. It is simple and assigns equal importance to all samples hence biased towards majority classes. Focal Loss extends BCE by down weighting easily classified samples and focus on learning and classifying hard samples. Improves sensitivity to minority classes and reduces dominance of vast number of easy negative samples [10]. Class-Balanced Loss re-weights each class based on its effective number of samples rather than its raw frequency, promoting more equitable learning across classes since minority classes receive proportionally higher emphasis during training [11].

3.3.5. Attention Mechanism

Attention mechanisms offer great push to CNNs by highlighting important features the model should pay closer attention relevant to the task at hand. In our work, we will explore Squeeze-and-Excitation (SE) Attention and Multi-Head Attention (MHA). SE attention performs channel-wise feature recalibration by modeling interdependencies between the feature channels using global context [12]. In retinal imaging context, SE attention highlights structures relevant to diagnostics(microaneurysms, exudates, drusen, hemorrhages, etc.) while suppressing less informative background features (noise, imaging artifacts etc.). MHA attention on the other hand, enables the model to focus on multiple feature subspaces simultaneously enabling the view of both local (microaneurysms, hemorrhages, etc.) and global (vessel topology, lesions spatial distribution, etc.) contexts

[13]. As a result, it provides improved modelling of spatially distributed pathological patterns across the samples.

3.3.6. Transformer Models

Transformer based models through their ability to capture long range dependencies have shown strong metrics in AI applications. Vision Transformer (ViT) treats samples as sequence of patches and therefore applies global self-attention capturing spatial distant retinal regions. This is instrumental for multi-label eye disease classification since abnormalities are not always localized. Swin transformer on the other hand utilizes shifted window self-attention making computation localized with cross window interaction. It has lower computational cost due to its hierarchical structure. Efficient in modeling localized patterns such as microaneurysms. ViT and Swin transformer are therefore great for multi-label retinal disease classification due to their complementary global and local feature perspectives.

3.3.7. GradCAM Explainable AI

To improve interpretability of AI models, tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) are utilized. In our case, Grad-CAM helps to analyze regions of the samples that contribute most and least to the predictions. It does this by generating heatmaps of each class in the dataset through the use of gradients of the predicted class and feature maps derived from final CNN layer. As a result, highly influential and less important spatial areas with the samples can be visualized.

3.3.8. Ensemble Learning Methods

Ensemble learning is the combination of two or more models to achieve better generalization as opposed to individual models. Different models may capture complementary features improving robustness of the model. Weighted averaging combines the predicted class probabilities from multiple models by taking a weighted sum of their outputs. The final probability for class c is given as below;

$$\widehat{p_c} = \sum_{i=1}^{N} w_i \; p_{i,c} \qquad (5)$$

$p_{i,c}$ probability that model i assigns to class c
w_i : weight of model i s
It preserves confidence information from individual models and allows stronger models to contribute more to the final prediction. It is also efficient and easy to implement.

4. RESULTS AND DISCUSSION

4.1. No Fine Tuning

Firstly, we evaluated the generalizability of pretrained EfficientNet and Resnet18 models by exploring their performance on the ODIR dataset. Here the model parameters are not updated since the backbone and classifier weights remain fixed. Our results are as below;

Table 4.1.1 EfficientNet Baseline Model

Metric\Disease	DR	Glaucoma	AMD
Precision	0.56	0.72	0.60
Recall	0.57	0.73	0.74
F1_Score	0.56	0.73	0.59
Kappa	0.12	0.46	0.25

Table 4.1.2 Resnet18 Baseline Model

Metric\Disease	DR	Glaucoma	AMD
Precision	0.52	0.71	0.63
Recall	0.52	0.68	0.75
F1_Score	0.50	0.69	0.65
Kappa	0.03	0.38	0.32

Baseline EfficientNet model shows strongest performance on Glaucoma with an F1-score of 0.73 and a Kappa of 0.46. However, it performs poorly on DR and AMD as shown by the low Kappa values 0.12 and 0.25.This highlights the limited generalizability of the pretrained weights without fine tuning. Similar pattern is observed in the Resnet18 baseline whose performance on AMD (F1= 0.65, Kappa = 0.32) is better followed by Glaucoma (F1= 0.69). However, it struggles with DR as shown by low Kappa of only 0.03. Overall, both models generalize reasonably well to Glaucoma but struggle a lot with DR and AMD conditions. This can be attributed to class imbalance. As a result, fine tuning is necessary. Onsite F1-score of 60.38 and 56.67 was achieved for EfficientNet and Resnet18 baseline models. This aligns with the offsite results highlighting that EfficientNet generalizes better to unseen data while ResNet18 remains less consistent.

4.2. Fine Tuning Classifier Only

In this setup, the convolution backbone is frozen and only the classifier layer is trained using the ODIR training dataset. The backbone role is fixed feature extractor while the classifier learns from the ODIR dataset.

Table 4.2.1 EfficientNet Classifier Tuned Model

Metric\Disease	DR	Glaucoma	AMD
Precision	0.88	0.49	0.24
Recall	0.59	0.76	0.82
F1_Score	0.71	0.59	0.37
Kappa	0.34	0.42	0.24

Table 4.2.2 Resnet18 Classifier Tuned Model

Metric\Disease	DR	Glaucoma	AMD
Precision	0.83	0.47	0.25
Recall	0.79	0.53	0.77
F1_Score	0.81	0.50	0.38
Kappa	0.41	0.32	0.27

The results in tables 4.2.1 and 4.2.2 shows that both models perform well on DR with different tradeoffs. EfficientNet has higher precision (0.88) but a lower recall (0.59) highlighting conservative predictions with fewer false positives but misses a lot of DR cases. On the other hand, Resnet18 achieves a higher F1-score(0.81) due to a more balanced precision-recall trade off.

EfficientNet shows a higher recall (0.76 and 0.82) for both Glaucoma and AMD respectively highlighting an improvement to minority classes. However, this comes at the cost of lower precision leading to a lower F1-score also. Resnet18 performance is relatively stable but still lower for both Glaucoma and AMD. This is further proved by onsite test results as F1-Score of 74.91 and 63.48 is achieved for EfficientNet and Resnet18 respectively. EfficientNet benefits more from classifier only fine tuning showing that its pretrained backbone is stronger and more generalizable. Classifier only fine tuning overally gave an improved performance while highlighting the limitation of fixed feature representations for minority disease detection.

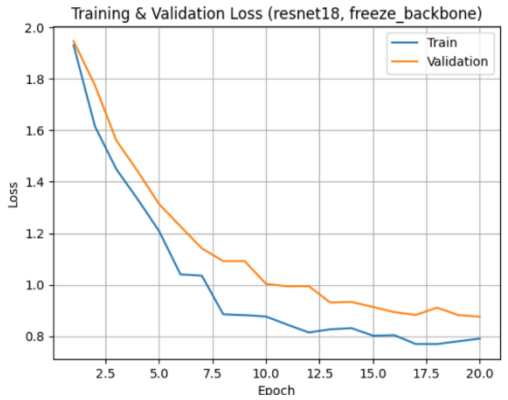

Figure 4.2.1 Training &Validation Loss Resnet18

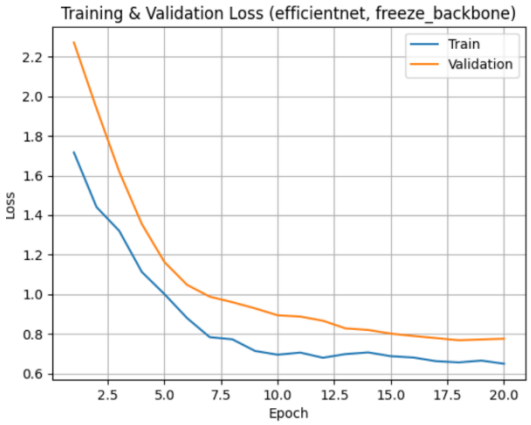

Figure 4.2.2 Training &Validation Loss EfficientNet

Both models show stable optimization highlighting low overfitting. EfficientNet however learns faster and reaches a lower loss value while Resnet18 convergence is slower and has higher loss plateau.

4.3. Full Fine Tuning

Here, both the backbone and classifier parameters are updated using the ODIR training dataset. This allows the model to extract meaningful low level and high-level features specific to our ODIR dataset. This leads to higher performance at the risk of overfitting.

Table 4.3.1 EfficientNet Full Fine-Tuned Model

Metric\Disease	DR	Glaucoma	AMD
Precision	0.91	0.63	0.31
Recall	0.70	0.85	0.73
F1_Score	0.79	0.72	0.44
Kappa	0.46	0.62	0.34

Table 4.3.2 Resnet18 Full Fine-Tuned Model

Metric\Disease	DR	Glaucoma	AMD
Precision	0.91	0.63	0.29
Recall	0.71	0.80	0.77
F1_Score	0.80	0.70	0.43
Kappa	0.47	0.59	0.32

As shown in Tables 4.3.1 and 4.3.2, full fine-tuning significantly improves performance across all diseases compared to classifier-only tuning. Both models exhibit a higher precision for DR (0.91) indicating a strong bias towards the dominant class. Substantial gains are observed for Glaucoma where EfficientNet achieves an F1-score of 0.72 while Resnet achieves 0.70. Thus, an improved sensitivity towards the minority classes. AMD performance though improves; it is still lower due to severe class imbalance. Onsite test set results are consistent with EfficientNet achieving an average F-score of 80.12 slightly outperforming Resnet18 which achieves 79.84. This confirms that full fine-tuning provides the best overall generalization as shown by the consistent gains across both offsite and onsite test sets.

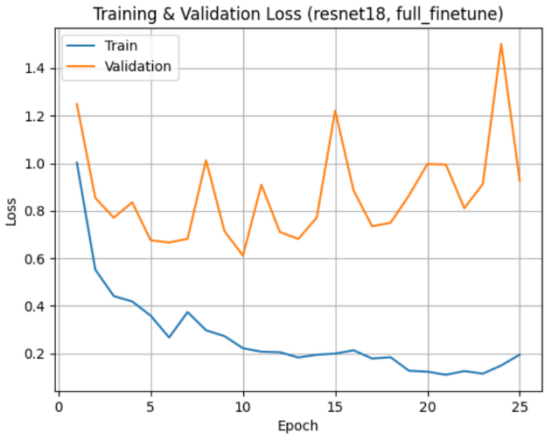

Figure 4.3.1 Training &Validation Loss Resnet18

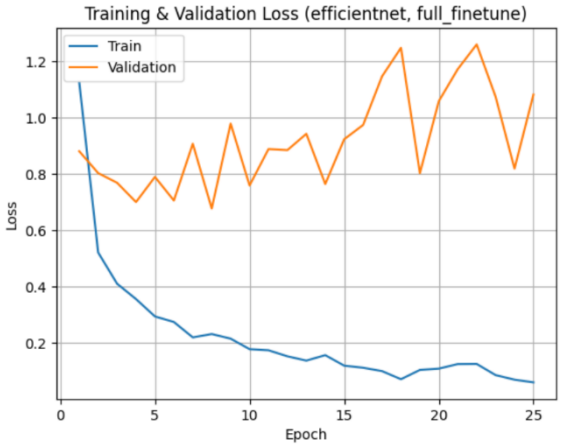

Figure 4.3.2 Training &Validation Loss Resnet18

Both models show a rapid training loss reduction highlighting that the model is learning faster and adapts to the ODIR dataset when fully fine-tuned. However, the validation loss fluctuates significantly, indicating a higher risk of overfitting compared to frozen-backbone training. Both show similar convergence strategies.

4.4. Focal Loss

To tackle class imbalance in our dataset, Focal Loss was applied to focus on hard to classify samples while down-weighting easy ones while training the model. EfficientNet achieves a better performance (F1-score 82.55) in comparison to Resnet18 (79.72).

Table 4.4.1 EfficientNet with Focal Loss

Metric\Disease	DR	Glaucoma	AMD
Precision	0.85	0.76	0.63
Recall	0.85	0.63	0.68
F1_Score	0.85	0.69	0.65
Kappa	0.50	0.60	0.61

As shown in table 4.4.1, EfficientNet achieves a strong and balanced results across the three conditions. An improvement of majority class DR F1-score (0.85) is observed while minority class AMD also attains a massive improvement from 0.44 to 0.65. The Kappa values for the minority class AMD also improve from 0.34 to 0.61 highlighting that Focal Loss effectively mitigates class imbalance. For Glaucoma, precision improves to 0.76 while recall decreases to 0.63 leading to a slight reduction on the F1-score 0.69. This shows the shift towards more conservative predictions reducing false positives but missing true Glaucoma cases.

Overally, the results highlight that focal loss enhances sensitivity to minority classes while maintaining strong generalization.

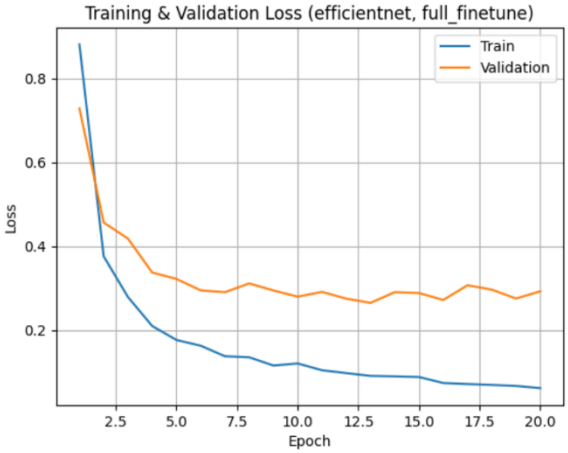

Figure 4.4.1 EfficientNet with Focal Loss.

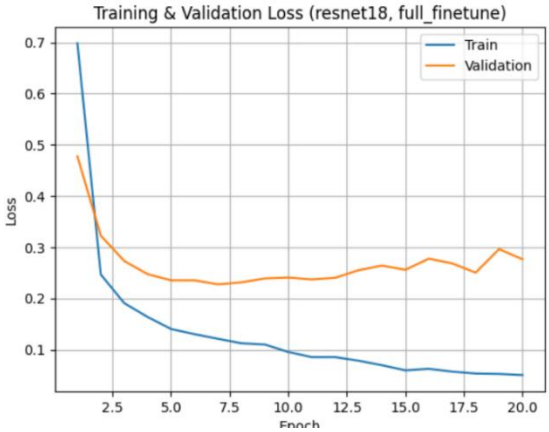

Figure 4.4.2 Resnet18 with Focal Loss.

Focal Loss stabilizes training and validation loss for both models, reducing overfitting compared to standard loss. EfficientNet converges faster and maintains a lower validation loss, while ResNet18 shows slightly higher validation fluctuations, indicating weaker generalization.

4.5. Class-Balanced Loss

Using Class-Balanced Loss, ResNet18 demonstrates the highest relative improvement compared to its full fine-tuned baseline. As shown in Table 4.5.1, ResNet18 achieves consistent gains across all diseases, particularly for minority classes. For Glaucoma, the F1-score increases from 0.70 to 0.75, accompanied by a substantial improvement in precision Similarly, AMD shows improvement in F1-score (0.43 → 0.65) highlighting the effectiveness of class re-weighting for severely underrepresented classes.

Table 4.5.1 Resnet18 with Class Balanced Loss

Metric\Disease	DR	Glaucoma	AMD
Precision	0.86	0.89	0.72
Recall	0.83	0.65	0.59
F1_Score	0.84	0.75	0.65
Kappa	0.50	0.69	0.61

On the onsite test set, Resnet18 improves from 79.84 to 80.13 with the use of class balanced loss.

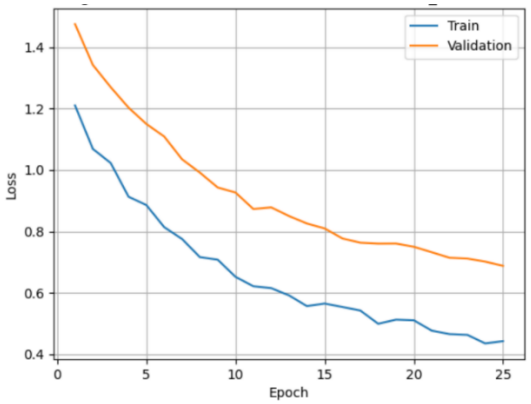

Figure 4.5.1 Efficient with Class Balanced Loss.

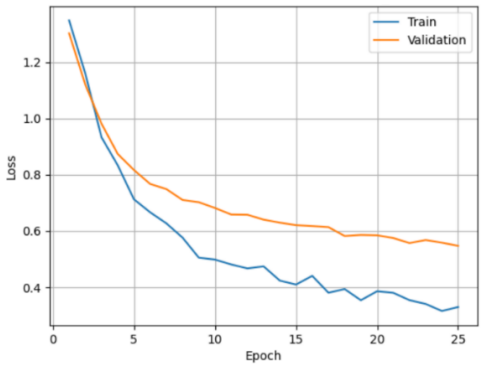

Figure 4.5.2 Resnet18 with Class Balanced Loss.

Both figures show smooth convergence with decreasing training and validation losses, indicating stable learning and improved handling of class imbalance using Class Balanced Loss.

4.6. Squeeze and Excitation Attention Mechanism

The incorporation of SE attention with full finetuning improved performance across all disease classes. DR achieves the highest F1-score (0.89) and recall (0.93) highlighting enhanced sensitivity. Other classes, Glaucoma and AMD F1-scores show an improvement increasing from 0.72 → 0.77 and 0.44 → 0.56 respectively, Although AMD is still the most challenging class.

Table 4.6.1 EfficientNet with SE Attention & BCE

Metric\Disease	DR	Glaucoma	AMD
Precision	0.85	0.80	0.50
Recall	0.93	0.73	0.63
F1_Score	0.89	0.77	0.56
Kappa	0.60	0.69	0.50

While focal and class-balanced losses were considered, standard BCE combined with threshold optimization yielded the most stable and consistent performance on the onsite tests with a F1-score of 82.099 better than Resnet18 80.691. Overall, SE attention led to consistent gains in F1-score and Cohen's Kappa. Thus, attention integration to model training improves robustness and feature discrimination more so for the minority classes. This is because of SE's ability to recalibrate channel responses using global view to stress on the fine features which are key to minority classes.

4.7. Multi-Head Attention (MHA)

Similarly, the application of MHA attention to the fully fined-tuned model leads to an improved offsite performance, particularly for the minority classes. The DR F1-score improves from 0.79 to 0.90 with a substantial recall improvement 0.70 to 0.92 highlighting an improvement in sensitivity.

Table 4.7.1 EfficientNet with MHA Attention & BCE

Metric\Disease	DR	Glaucoma	AMD
Precision	0.88	0.73	0.72
Recall	0.92	0.65	0.59
F1_Score	0.90	0.69	0.65
Kappa	0.64	0.59	0.61

The AMD minority class gains significantly from MHA with its F1-score improving from 0.44 → 0.65 and Cohen's Kappa from 0.34 → 0.61. Glaucoma performance remains similar to baseline in primary metric (F1-score), suggesting that MHA focuses on minority class as opposed to uniformly improving all classes. The onsite test set gives an F1-score of 80.09 comparable to the baseline but does not show superior generalization to unseen data. This behavior may be attributed to factors such as loss function choice, threshold applied, and the limited size of the training dataset.

4.8. ViT and Swin Transformer Models

Fully fine-tuned ViT with Focal Loss function achieved an onsite F1-score of 0.8165 outperforming Swin's 0.7634. This highlights that ViT's global self-attention is more effective in capturing spatial retinal disease patterns. Nevertheless, both transformers performed well comparatively to the baseline CNN architecture. While transformers have advanced attention details, their gains are influenced by dataset size, class imbalance and tuning strategies employed.

4.9. GradCAM Explainable AI

Applying Gram-CAM to our samples show that the model mainly focuses on clinically relevant areas of the retina as opposed to background areas which could be possibly be noise from visual effects such as glare. High activations are located around the center of the retina which is associated with retinal disease manifestation. On the other hand, areas with low activation are anatomically less informative regions highlighting that the model focuses less on irrelevant visual cues as expected even with manual diagnostics. Grad-CAM in our case gave a qualitative validation that the model

focuses on clinically relevant regions as opposed to suboptimal regions. Figure 4.8.1, shows Grad-CAM visualization for retinal disease prediction and 4.8.2, the input sample to Grad-CAM. Warmer colors indicate regions with higher contribution to the model's prediction, highlighting anatomically relevant retinal areas.

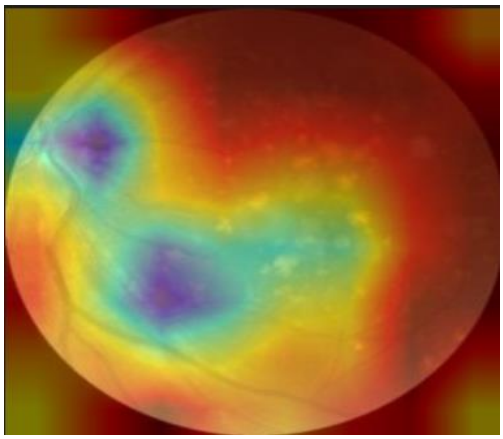

Figure 4.8.1: Grad-CAM visualization for retinal disease prediction.

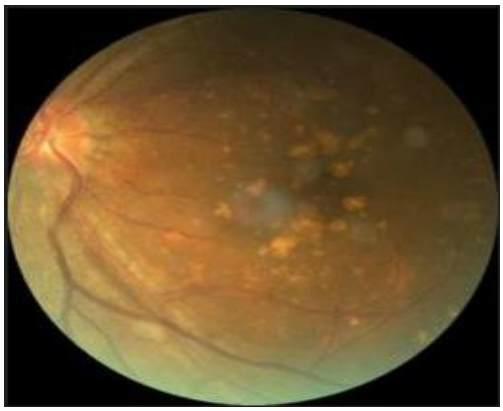

Figure 4.8.2: Original retinal sample used as Input for Grad-CAM.

4.10. Weighted Ensemble Learning

Several combinations of loss functions and attention mechanisms were used to analyze the impact of ensemble learning on model performance and generalization. Squeeze and excitation with focal loss gave the best comparative results.

Table 4.8.1 Ensemble with SE & Focal Loss

Metric\Disease	DR	Glaucoma	AMD
Precision	0.85	0.70	0.70
Recall	0.93	0.82	0.55
F1 Score	0.89	0.75	0.62
Kappa	0.59	0.67	0.57

The use of weighted ratio 0.55 and 0.45 for EfficientNet and Resnet18 led to an onsite F1-score of 79.15. Although this performance is lower than EfficientNet score of 79.7, it exceeds the Resnet18 score of 77.41. This indicates that while the ensemble does not surpass the strongest individual model on the onsite dataset, it provides a more balanced and robust performance by combining complementary model strengths as shown by the improvement on Resnet18 performance.

5. CONCLUSION

This work has shown that baselines of CNN models can be surpassed by careful architectural and design strategies as opposed to brute model scaling. Applying pretrained EfficientNet and ResNet18 baselines to multi-label retinal disease classifications we were able to show the progressive gain by fine tuning the model architecture, applying loss functions to handle class imbalance and attention mechanisms for improved learning. Fully fine-tuned model with Focal Loss or Class-Balanced Loss improved the minority class classification while attention mechanism especially Squeeze and Excitation attention with Binary Cross Entropy loss gave the highest overall F1-score on the onsite test. Attention integration with loss functions such as Focal Loss and Class-Balanced Loss gave lesser comparative results highlighting the complexity of gains in modeling. Transformer models (ViT and Swin) also gave competitive results though the performance gains were minimal due to the dataset size and imbalance inherent.

The use of ensemble learning (weighted) improved model robustness of the weaker model at the cost of the strongest individual model highlighting marginal gains when model diversity is constrained. Finally, the adoption of Grad-CAM helped to validate qualitatively how the baselines focus on clinically relevant parts of the fundus as opposed to noise regions improving trust in predictions. Overall, this work presents a crafted methodology for building and fine tuning robust and reliable multi-label medical imaging system by emphasizing the need of loss design and attention strategies, architectural change and ensemble learning to surpass the baselines. Future work will explore the integration of Grad-CAM into the training process to improve learning and generalization.

6. REFERENCES

[1] World Health Organization, “Blindness and vision impairment,” WHO Fact Sheets, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] M. J. Burton, J. Ramke, A. P. Marques, R. R. Bourne, N. Congdon, I. Jones, et al., “The Lancet Global Health Commission on Global Eye Health: vision beyond 2020,” Lancet Global Health, vol. 9, no. 4, pp. e489–e551, 2021.

[3] World Bank, “Physicians (per 1,000 people),” Global Health Workforce Statistics, WHO, OECD, 2025. [Online]. Available: https://data.worldbank.org/indicator/SH.MED.PHYS.ZS

[4] H. Fang, F. Li, H. Fu, *et al.*, “Adam challenge: Detecting age-related macular degeneration from fundus images,” *IEEE Trans. Med. Imaging*, vol. 41, no. 10, pp. 2828–2847, 2022.

[5] H. Fang, F. Li, J. Wu, *et al.*, “Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening,” *arXiv preprint* arXiv:2202.08994, 2022.

[6] M. Karthik, M. Maggie, and S. Dane, “APTOS 2019 Blindness Detection,” Kaggle, 2019. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[7] A. Müller, I. Soto-Rey, and F. Kramer, “Towards a guideline for evaluation metrics in medical image segmentation,” BMC Research Notes, vol. 15, no. 210, Jun. 2022.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” Adv. Neural Inf. Process. Syst., vol. 27, pp. 3320–3328, 2014.

[9] N. Tajbakhsh et al., "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299-1312, May 2016.

[10] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007.

[11] Y. Cui, M. Jia, T. -Y. Lin, Y. Song and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9260-9269.

[12] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.