Diabetic Retinopathy Classification – Final Report

EEL4759: Digital Image Processing

Vineet Kumar

April 21, 2026

1 Abstract

This report covers automated severity grading of diabetic retinopathy (DR) from retinal images using deep convolutional neural networks. Three pretrained architectures (ResNet-50, EfficientNet-B0, and ViT-B/16) were fine-tuned on the Kaggle "Diabetic Retinopathy Resized Arranged" dataset and evaluated on a five-class severity classification task (Healthy through Proliferative DR). Each model was trained with and without CLAHE (Contrast Limited Adaptive Histogram Equalization) preprocessing to assess its effect on classification performance. Class imbalance was addressed via sqrt-inverse-frequency weighted sampling and Focal Loss. At 384×384 pixel resolution, ResNet-50 and EfficientNet-B0 achieved 0.80 overall accuracy and weighted F1-scores of 0.77 and 0.78 respectively. ViT-B/16 was constrained to 224×224 and achieved 0.69 accuracy. CLAHE improved detection of minority classes (particularly Mild NPDR) at a small cost to overall accuracy. Mild NPDR remained the hardest class to classify across all experiments, with F1-scores no higher than 0.17 due to its visual similarity to healthy retinal images.

2 Introduction

2.1 Diabetic Retinopathy

Diabetic retinopathy is a progressive eye disease caused by diabetes that is one of the leading causes of preventable vision loss worldwide. It develops when high blood sugar damages the retinal blood vessels, causing leakage, abnormal blood vessel growth, and eventually retinal detachment if untreated. The disease is classified into five severity grades: no retinopathy (Healthy), Mild Non-Proliferative DR (NPDR), Moderate NPDR, Severe NPDR, and Proliferative DR. Early detection through systematic retinal screening can prevent up to 90% of severe vision loss cases. Manual grading by trained specialists is accurate but expensive and slow, making automated image classification useful for large-scale screening programs.

2.2 Convolutional Neural Networks for Retinal Image Classification

Convolutional neural networks (CNNs) have become the dominant approach for medical image classification tasks. Rather than hand-crafting features, CNNs learn hierarchical representations (low-level edges and textures in early layers, progressively more abstract features in deeper layers) directly from training data. For retinal imaging, pretrained ImageNet models (transfer learning) are effective: the low-level filters learned on natural images transfer well to retinal images, and fine-tuning on DR data allows the network to specialize to lesion-specific features such as microaneurysms, hard exudates, and hemorrhages that distinguish severity grades. This project compares three architectures spanning different design philosophies: ResNet-50 (residual connections), EfficientNet-B0 (compound scaling), and ViT-B/16 (pure self-attention on image patches).

2.3 Existing Model Performance

The Kaggle Diabetic Retinopathy detection competitions have established rough baselines for this type of task. On five-class grading datasets similar to the one used here, top-performing single models typically achieve quadratic weighted kappa scores in the 0.82–0.86 range using heavily tuned ensembles, test-time augmentation, and competition-grade preprocessing. More comparable single-model baselines reported on Kaggle notebooks for the "Diabetic Retinopathy Resized Arranged" dataset (224×224) report accuracies of roughly 0.60–0.73 and kappa values around 0.50–0.58 for standard fine-tuned CNNs, which is consistent with the results obtained in this project's 224px training run.

3 Methodology

3.1 Dataset and Class Distribution

The dataset used is "Diabetic Retinopathy Resized Arranged" [1], containing 35,126 retinal images organized into five class-labeled folders. The images are JPEG files at approximately 1024×768 pixel resolution. The dataset was split into 70% training, 15% validation, and 15% test sets using stratified sampling to preserve the class distribution across all splits (scikit-learn's `train_test_split` with `stratify`, seed 42).

Table 1: Dataset class distribution (test split, 5,269 images).

Class	Grade	Test Count	Test %
0	Healthy	3,872	73.5%
1	Mild NPDR	366	6.9%
2	Moderate NPDR	794	15.1%
3	Severe NPDR	131	2.5%
4	Proliferative DR	106	2.0%

The dataset is severely imbalanced: the Healthy class accounts for nearly 74% of all samples, while Severe NPDR and Proliferative DR together comprise only 4.5%.

3.2 Models

Three ImageNet-pretrained architectures were fine-tuned for 5-class DR severity classification. In all cases the original classification head was replaced with a linear layer mapping to 5 outputs, and all backbone weights were unfrozen for full fine-tuning.

- ResNet-50 [3]: Deep residual network with skip connections. The final fully connected layer ($2048 \rightarrow 5$) was replaced. Input: 384×384.
- EfficientNet-B0 [5]: Compound-scaled network optimizing width, depth, and resolution simultaneously. The classifier head (`Dropout(0.2)` + Linear $1280 \rightarrow 5$) was replaced. Input: 384×384.
- ViT-B/16 [2]: Vision Transformer that splits the image into 16×16 pixel patches and processes them with multi-head self-attention. The projection head ($768 \rightarrow 5$) was replaced. Input: 224×224 (PyTorch's ViT-B/16 implementation only supports 224×224 input; higher resolutions are not supported without custom modifications to the model).

3.3 Preprocessing and Augmentation

Images were resized from their native ˜1024×768 resolution to the model's target input resolution (384×384 for ResNet-50 and EfficientNet-B0; 224×224 for ViT-B/16) as the first preprocessing step.

CLAHE (Contrast Limited Adaptive Histogram Equalization) [6] was applied as an optional subsequent preprocessing step. The transform converts the image from RGB to LAB color space, applies `cv2.createCLAHE` (clip limit 2.0, tile grid 8×8) to the L (lightness) channel only, and converts back to RGB. Operating on the lightness channel alone enhances local contrast in retinal structures such as vessels and lesions without altering hue or saturation.

The training augmentation pipeline (applied after CLAHE if enabled):

1. `RandomResizedCrop` to target resolution, scale 0.8–1.0
2. `RandomHorizontalFlip` (p=0.5), `RandomVerticalFlip` (p=0.5)
3. =RandomRotation=($\pm 30°$)
4. `ColorJitter` (brightness 0.3, contrast 0.3, saturation 0.2, hue 0.02)
5. `GaussianBlur` (kernel 3, σ 0.1–1.0)
6. Normalize to ImageNet mean/std

Validation and test images were resized to the target resolution deterministically (no random crop) and normalized identically.

3.4 Class Imbalance Handling

Two complementary mechanisms addressed the severe class imbalance:

- WeightedRandomSampler: Each training sample was assigned a weight proportional to $1/\sqrt{n_c}$ where n_c is the count of its class. The square-root weighting provides a moderate upsampling of minority classes without completely drowning the majority-class signal (which full inverse-frequency weighting was found to do).
- Focal Loss [4]: The loss function $\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$ with $\gamma = 2$ down-weights easy, well-classified examples and focuses gradient updates on hard misclassifications, which helps with minority classes where the model initially predicts low confidence.

3.5 Training Configuration

Table 2: Hyperparameters used for all experiments.

Hyperparameter	Value
Optimizer	AdamW
Head learning rate	1e-4
Backbone learning rate	1e-5 (10× lower)
Weight decay	1e-4
Batch size	256
Max epochs	50
Early stopping	patience 10, monitor val macro F1
LR warmup	3 epochs linear ($0.1\times \rightarrow 1\times$)
LR schedule	Cosine annealing after warmup
Gradient clipping	max norm 1.0
Mixed precision	`torch.amp.autocast` + GradScaler
Hardware	2× AMD Radeon RX 7900 XTX (ROCm)
Random seed	42

A discriminative learning rate was used: the pretrained backbone was trained at 1e-5 while the newly initialized classification head was trained at the full 1e-4. This prevents the pretrained features from being overwritten too quickly in early epochs.

3.6 Evaluation Metrics

Each trained model was evaluated on the held-out test set using:

- Accuracy: fraction of correctly classified samples
- Weighted F1: F1 averaged over classes weighted by support; reflects overall performance on the imbalanced distribution

- Macro F1: F1 averaged equally over all 5 classes; better reflects performance on minority classes
- Quadratic-weighted Cohen's Kappa: measures inter-rater agreement beyond chance; the quadratic weighting penalizes predictions further from the true label more heavily, which is appropriate for the ordinal DR severity scale

4 Results

4.1 Summary of Test-Set Performance

Table 3: Test-set metrics for all six experiments at 384×384 (ResNet-50, EfficientNet-B0) and 224×224 (ViT-B/16).

Model	CLAHE	Accuracy	Weighted F1	Macro F1
ResNet-50	No	0.80	0.77	0.51
ResNet-50	Yes	0.78	0.77	0.53
EfficientNet-B0	No	0.80	0.78	0.52
EfficientNet-B0	Yes	0.78	0.76	0.51
ViT-B/16	No	0.68	0.69	0.47
ViT-B/16	Yes	0.69	0.71	0.49

4.2 Per-Class F1 Scores

Table 4: Per-class F1-score for all six experiments.

Model / CLAHE	Healthy	Mild NPDR	Moderate NPDR	Severe NPDR	Proliferative DR
ResNet-50 / No	0.90	0.09	0.57	0.50	0.52
ResNet-50 / Yes	0.89	0.17	0.53	0.45	0.61
EfficientNet-B0 / No	0.90	0.10	0.57	0.45	0.58
EfficientNet-B0 / Yes	0.88	0.11	0.54	0.45	0.60
ViT-B/16 / No	0.81	0.14	0.43	0.43	0.51
ViT-B/16 / Yes	0.83	0.17	0.45	0.45	0.56

4.3 Effect of Input Resolution (224px vs 384px)

The experiments were run twice: an initial run at 224×224 and a subsequent run at 384×384 for ResNet-50 and EfficientNet-B0. ViT-B/16 was kept at 224×224 in both runs. The resolution increase produced a substantial improvement for the CNN models.

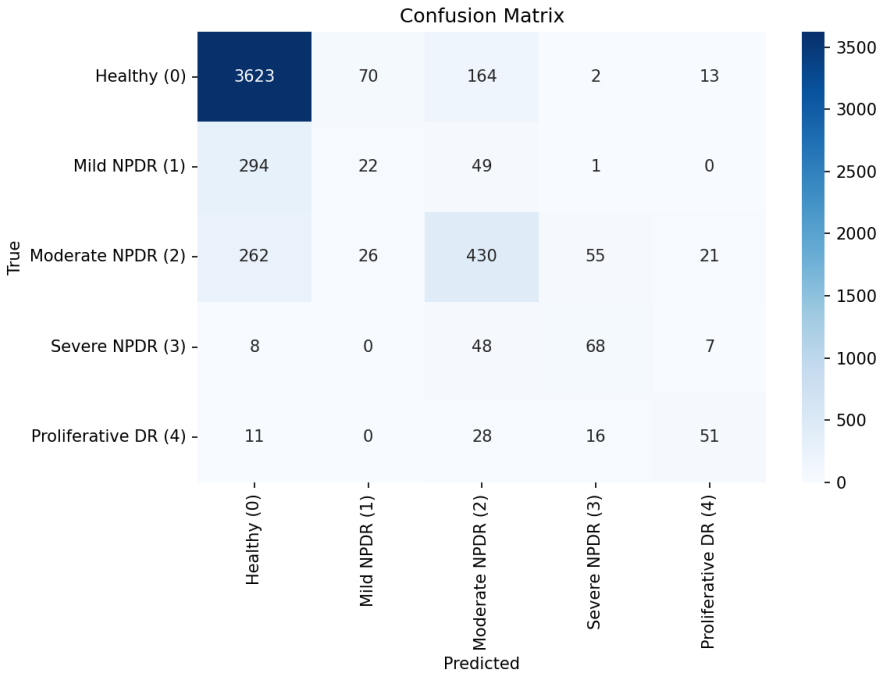
Table 5: Comparison of 224px (initial run) vs 384px (final run) test-set results for CNN models. ViT results shown for reference; both runs used 224px.

Model	CLAHE	Acc (224px)	W-F1 (224px)	M-F1 (224px)	Kappa (224px)	Acc (384px)	W-F1 (384px)	M-F1 (384px)
ResNet-50	No	0.61	0.66	0.46	0.53	0.80	0.77	0.51
ResNet-50	Yes	0.61	0.65	0.44	0.52	0.78	0.77	0.53
EfficientNet-B0	No	0.61	0.65	0.46	0.55	0.80	0.78	0.52
EfficientNet-B0	Yes	0.62	0.66	0.45	0.54	0.78	0.76	0.51
ViT-B/16	No	0.68	0.69	0.47	0.53	0.68	0.69	0.47
ViT-B/16	Yes	0.67	0.69	0.48	0.56	0.69	0.71	0.49

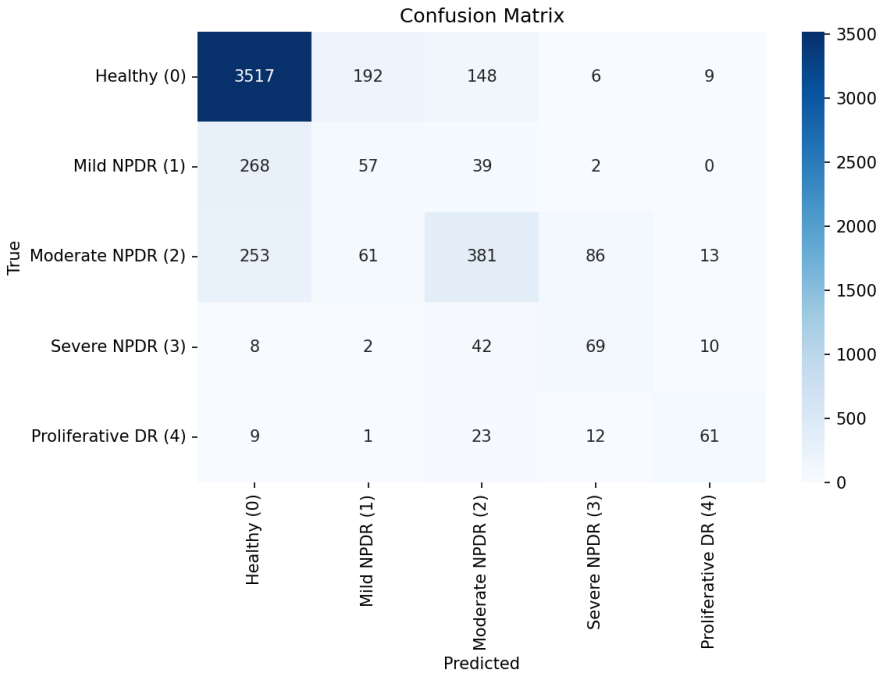
Accuracy for ResNet-50 and EfficientNet-B0 improved by approximately 19 percentage points (0.61 $\rightarrow$ 0.80) with the resolution increase. The improvement is consistent across both CLAHE settings and is attributed to the finer retinal microstructure (microaneurysms, hemorrhage dots, exudates) that is present at 384px but lost at 224px.

4.4 Confusion Matrices

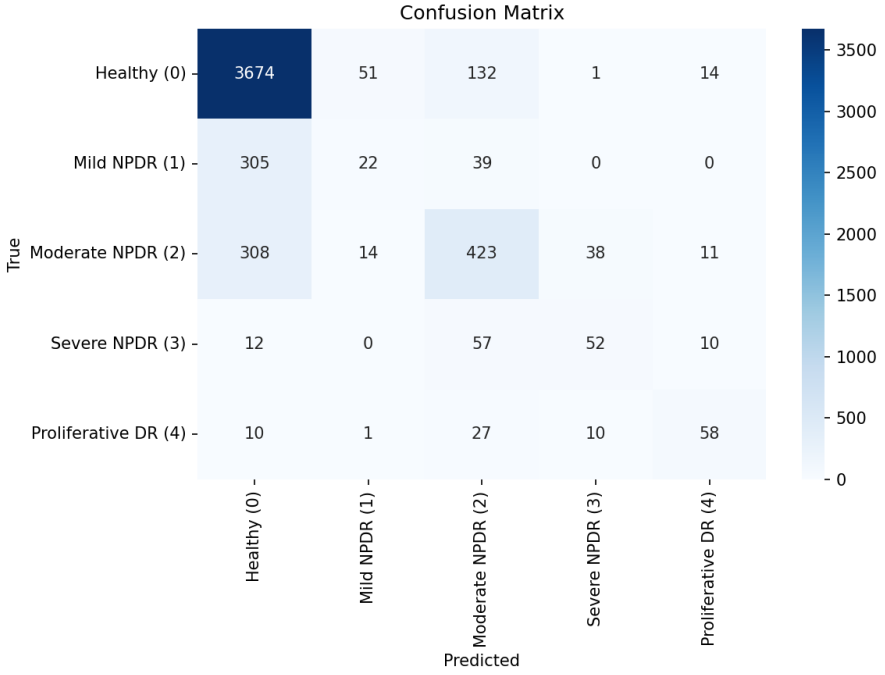
4.4.1 ResNet-50 without CLAHE



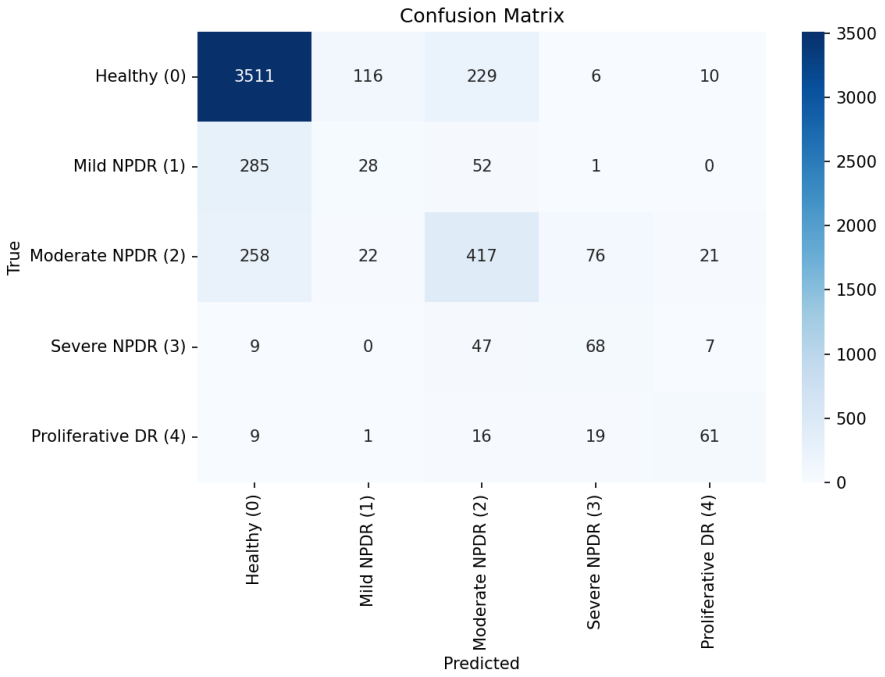
4.4.2 ResNet-50 with CLAHE



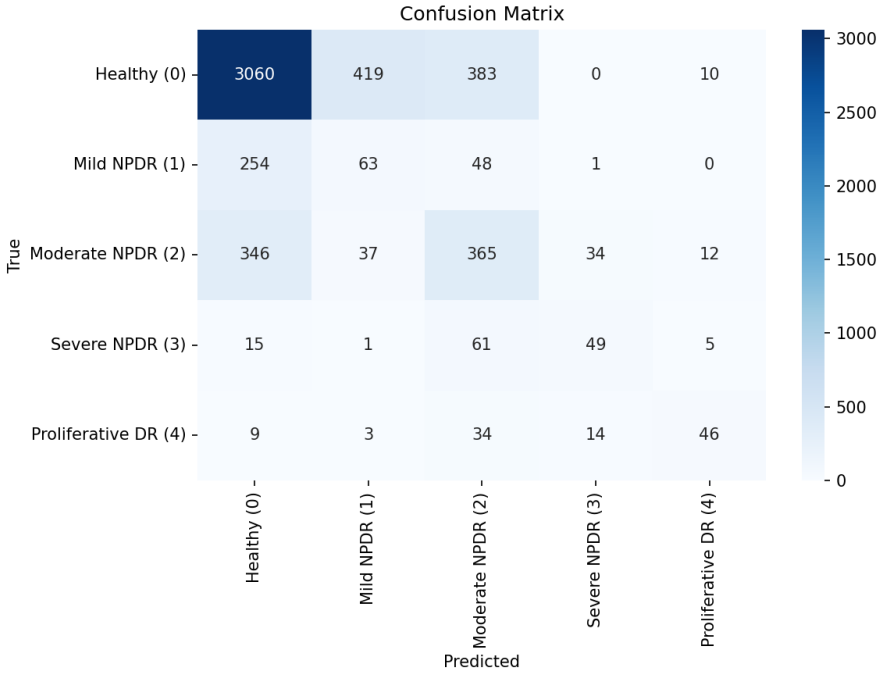
4.4.3 EfficientNet-B0 without CLAHE



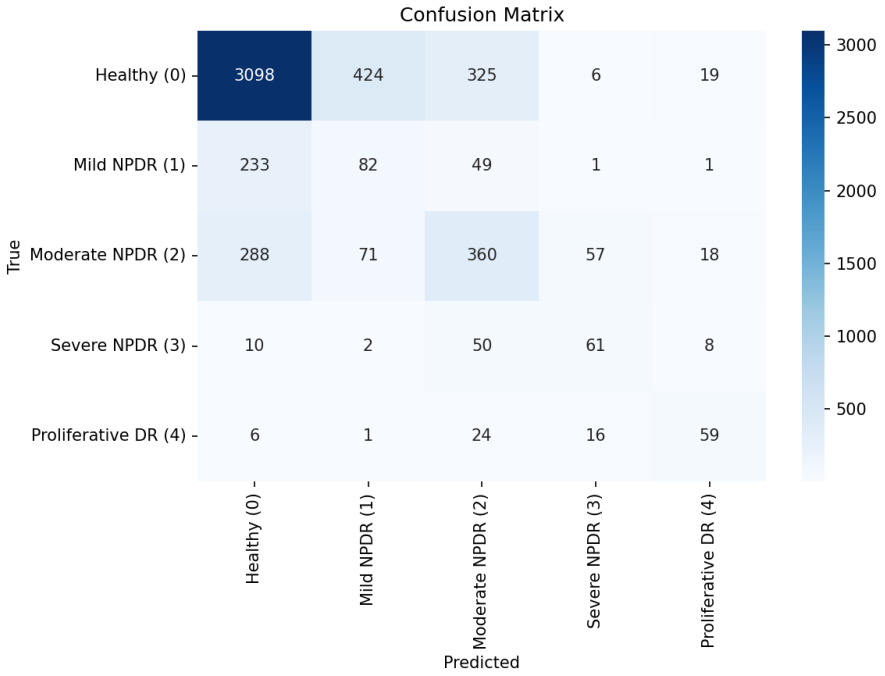
4.4.4 EfficientNet-B0 with CLAHE



4.4.5 ViT-B/16 without CLAHE

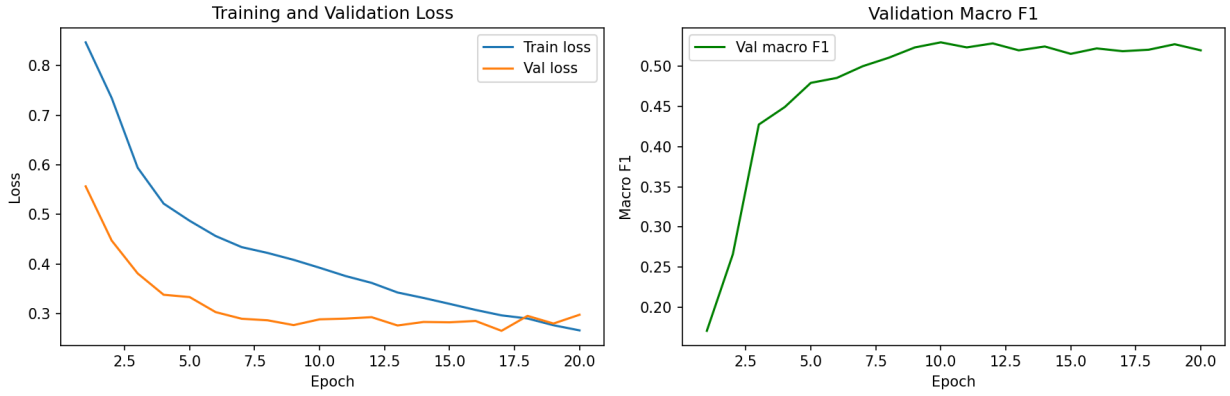


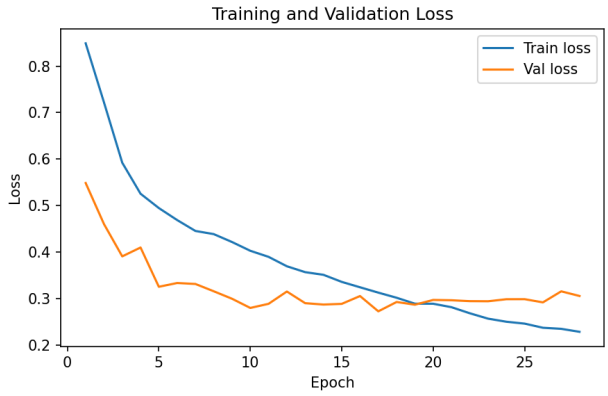
4.4.6 ViT-B/16 with CLAHE

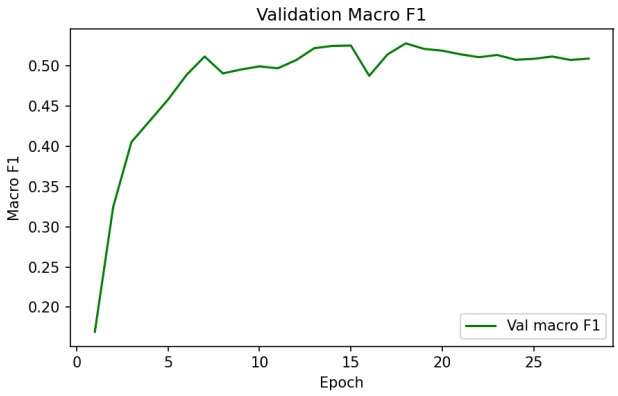


4.5 Training Curves

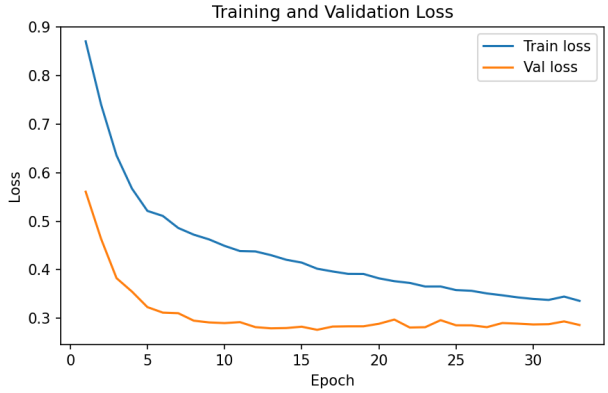
4.5.1 ResNet-50

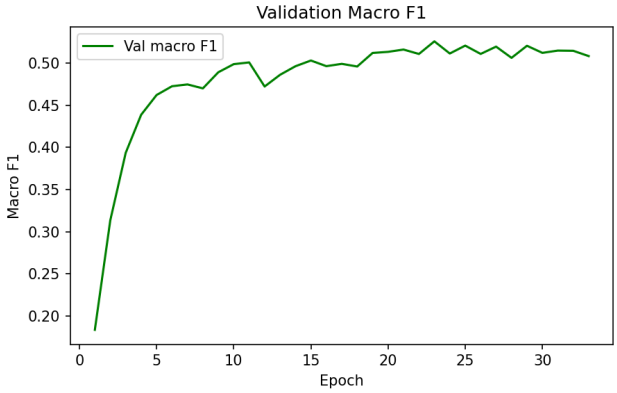


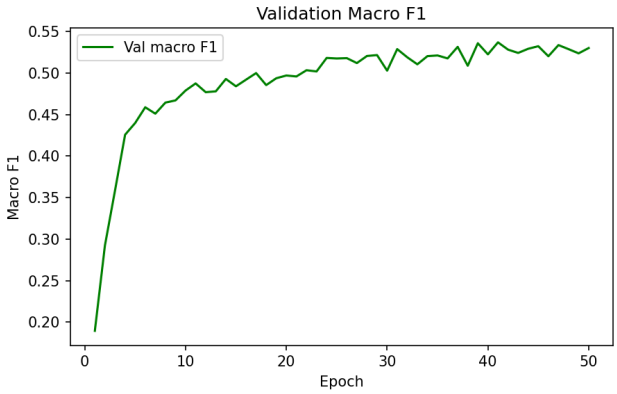




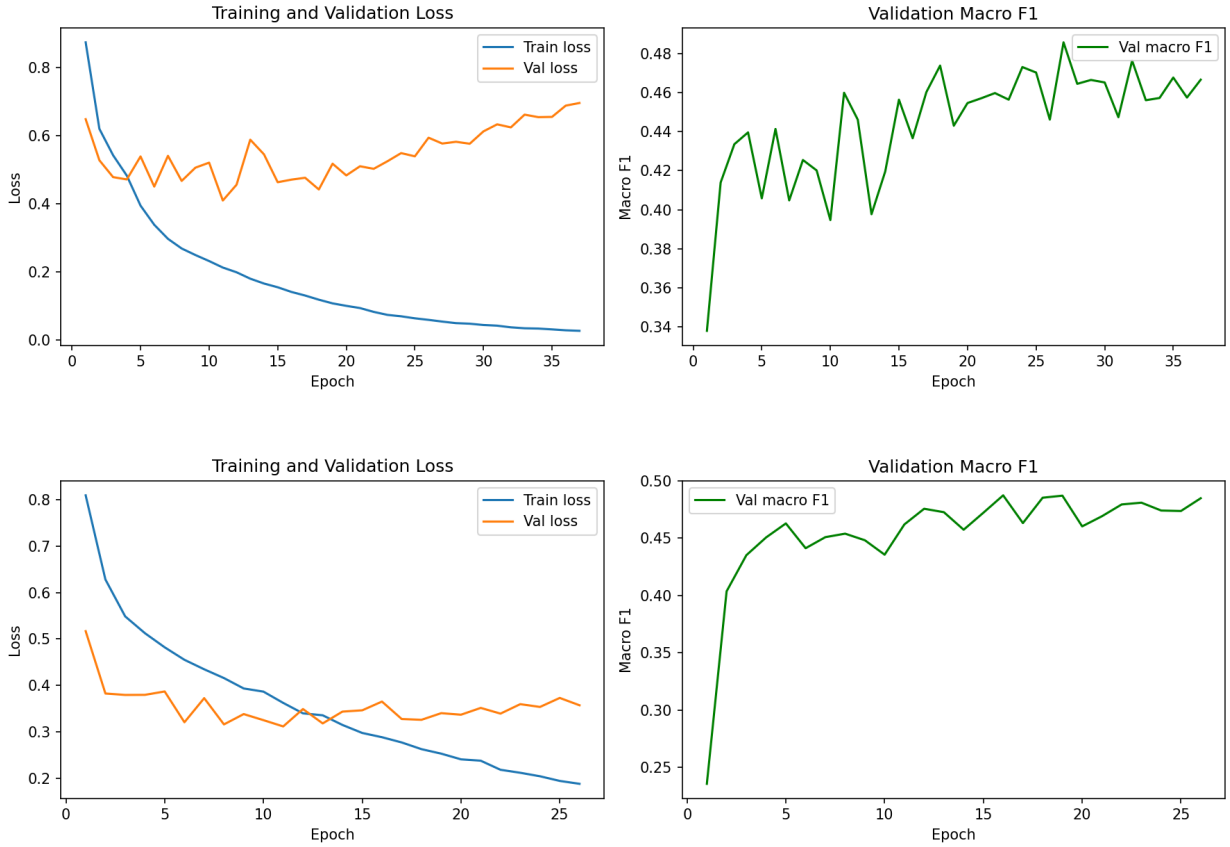
4.5.2 EfficientNet-B0







4.5.3 ViT-B/16



5 Discussion

5.1 Resolution Is the Dominant Factor for CNN Models

The largest improvement was the ˜19 percentage point accuracy jump for ResNet-50 and EfficientNet-B0 when input resolution was increased from 224×224 to 384×384. Retinal images contain small features used for diagnosis (microaneurysms, dot-like red lesions as small as 10–20μm, dot hemorrhages, and hard exudates) that take up only a few pixels at 224px. At 384px these features are resolved clearly enough for the network to learn discriminative filters for them. The resolution improvement had no effect on ViT-B/16 because PyTorch's ViT-B/16 implementation only supports 224×224 input and cannot be run at higher resolutions without custom modifications to the model.

5.2 CLAHE Trades Accuracy for Minority-Class Recall

Applying CLAHE consistently reduced overall accuracy by approximately 2 percentage points for the CNN models (e.g., ResNet-50: 0.80 → 0.78) while improving detection of minority classes. For ResNet-50, CLAHE nearly doubled the Mild NPDR F1-score from 0.09 to 0.17, and improved Proliferative DR F1 from 0.52 to 0.61. This happens because

CLAHE enhances the local contrast of subtle lesions, making minority-class features more distinguishable, but the enhanced contrast can also introduce artifacts that disrupt features the model relied on for the majority Healthy class.

For ViT-B/16, CLAHE produced a small but consistent improvement across most classes, and the Mild NPDR recall improved from 0.17 to 0.22. This suggests that CLAHE is more beneficial for ViT, possibly because the transformer's attention mechanism can exploit the enhanced contrast across longer-range spatial dependencies.

Whether CLAHE is desirable depends on the application: maximizing overall accuracy favors CLAHE-off, while maximizing detection of the higher-risk advanced DR grades favors CLAHE-on. For a clinical screening tool the latter is generally preferable.

5.3 Mild NPDR Remains Universally Difficult

Mild NPDR (class 1) had the lowest F1-score in every experiment, ranging from 0.09 to 0.17. This class is defined by only microaneurysms with no other lesions, a very subtle change from the healthy retina. The confusion matrices confirm that the vast majority of Mild NPDR predictions are misclassified as Healthy. The visual difference is small, the class is heavily underrepresented (6.9% of the dataset), and the class boundaries are subjective even for trained graders. Despite weighted sampling and Focal Loss, the network does not reliably learn to distinguish these cases. A possible solution to this is to train with a classification model that supports higher resolutions such as the modified Hi-ResNet, where I hypothesize that the small lesions in class 1 get obscured or even essentially erased when downscaling the images during the dataset preparation for use as input to the models.

5.4 ViT-B/16 Underperforms at 224px

ViT-B/16 achieved lower accuracy (0.68–0.69) compared to ResNet-50 and EfficientNet-B0 (0.78–0.80). This is mainly due to the resolution constraint: ViT-B/16 uses 16×16 pixel patches, so at 224px each patch covers a 16×16 area, which is large enough to subsume entire microaneurysms. The training curves reveal severe overfitting: the training loss approaches zero while validation loss increases after ˜10–15 epochs. ViT-B/16 likely requires either a higher input resolution (which demands interpolated positional embeddings and careful fine-tuning), a larger dataset, or stronger regularization to generalize well on medical images.

6 Conclusion

Pretrained CNN architectures achieved strong diabetic retinopathy grading performance (accuracy ˜0.80, weighted F1 ˜0.77–0.78) when fine-tuned at sufficient resolution (384×384). Input resolution was the most impactful factor, with a ˜19 percentage point accuracy improvement for ResNet-50 and EfficientNet-B0 when resolution was increased from 224 to 384 pixels. CLAHE preprocessing improved minority class detection at a small overall accuracy cost and is recommended for clinical settings where detecting advanced DR grades is the priority. EfficientNet-B0 achieved the best weighted F1 (0.78) without CLAHE, while

ResNet-50 with CLAHE achieved the best macro F1 (0.53), indicating slightly more balanced performance across classes. ViT-B/16 was constrained by its fixed-resolution patch embedding and underperformed the CNN models.

Mild NPDR remains the hardest class: its visual distinction from a healthy retina is subtle and the class is severely underrepresented. Future work could address this with: dedicated data augmentation for lesion synthesis, higher-resolution ViT variants (ViT-L with interpolated positional embeddings), pre-training on larger retinal image datasets, or model ensembles.

7 Appendix

The repository will be available at `https://git.vineetk.net/eel4759_finalproject_classification/`.

References

[1] Amanneo. *Diabetic Retinopathy Resized Arranged*. 2022. URL: `https://www.kaggle.com/datasets/amanneo/diabetic-retinopathy-resized-arranged`.

[2] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*. 2021.

[3] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[4] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988.

[5] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019, pp. 6105–6114.

[6] Karel Zuiderveld. "Contrast Limited Adaptive Histogram Equalization". In: *Graphics Gems IV*. Ed. by Paul S. Heckbert. Academic Press, 1994, pp. 474–485.