Multi-Task ViT and CNN Architectures for Diabetic Retinopathy Severity Classification

Ata Güneş
Electrical and Electronics Engineering
Izmir Democracy University
Izmir, Türkiye
ata_gn@hotmail.com

Esra Cesur
Electrical and Electronics Engineering
Izmir Democracy University
Izmir, Türkiye
esracesur20@gmail.com

***Abstract*—This report addresses the crucial task of classifying Diabetic Retinopathy (DR) severity from retinal fundus images, navigating the challenge of class imbalance using deep learning. Our research involved robust image preprocessing and extensive data augmentation to enhance critical features. We explored five distinct deep learning architectures, ranging from single-stage EfficientNets to advanced two-stage Vision Transformer (ViT) models, incorporating techniques like multi-task learning and Focal Loss. Evaluated using Cohen's Kappa, our most refined two-stage ViT model (Model 5) achieved the highest Kaggle accuracy of 0.86075, securing 2nd place in the competition. These results highlight the effectiveness of advanced preprocessing, tailored loss functions, multi-task learning, and powerful Transformer architectures in achieving accurate DR diagnosis.**

***Index Terms*—Diabetic Retinopathy, Vision Transformer (ViT), Medical Imaging, Multi-task Learning, Deep Learning**

Colab Notebook: https://colab.research.google.com/drive/1jKgU-lRVS0r3xyfWo-FZ18Kfub5JIQ6T?usp=sharing

I. Introduction

Diabetic Retinopathy (DR) stands as a leading cause of preventable blindness worldwide, a severe complication arising from long-term diabetes. Its early detection and accurate classification are paramount for timely intervention and preventing irreversible vision loss. The progression of DR is typically categorized into distinct stages, ranging from "No DR" to severe forms like "Proliferative DR," each characterized by specific lesions and abnormalities observable in retinal fundus images.

As **Fig. 1** illustrates, the visual manifestations of DR evolve with severity. Healthy retinas (No DR) exhibit clear blood vessels and optic disc. However, as the disease progresses through stages like Mild, Moderate, Severe Non-Proliferative Diabetic Retinopathy (NPDR), and eventually Proliferative DR, characteristic signs emerge. These include micro-aneurysms, hemorrhages, soft or hard exudates, intra-retinal microvascular abnormalities (IRMAs), and neovascularizations. The ability to precisely identify and categorize these subtle yet critical features from fundus photographs is central to effective diagnosis and management.

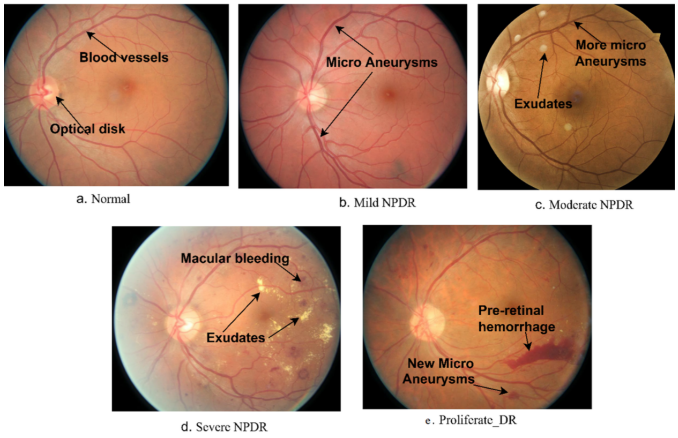

Fig. 1: Diabetic Retinopathy Stages [1]

This report details our development of advanced deep learning models to automate the classification of DR severity from these complex retinal images. Our approach aims to not only detect DR but also to accurately discern its various stages, providing a crucial tool for ophthalmologists.

II. Previous studies

A comprehensive review highlights key trends and methodologies, summarized below:

Che et al. (2022) [2] reviewed 83 methods; effective preprocessing included CLAHE, circular masking, and green channel extraction. Ensemble models improved accuracy by 2-5% over single models.

Alyoubi et al. (2022) [3] compared 7 CNNs on Kaggle EyePACS; EfficientNetB3 achieved 96.62% accuracy with 43% fewer parameters than ResNet50. ImageNet pre-training accelerated convergence by 60% and improved accuracy by 3-4%.

Pratiwi et al. (2023) [4] proposed a preprocessing pipeline (green channel, CLAHE, gamma correction, circular ROI, augmentation). This improved ResNet50 accuracy by 7.3% and EfficientNetB4 by 5.1%, with optimal Adam optimizer settings (1e-4 LR, batch 16).

Fang et al. (2024) [5] introduced a hybrid CNN-ViT architecture (EfficientNetB0 feature extractor, custom transformer, hybrid loss). Achieved 94.8% accuracy on multi-institutional datasets, showing high sensitivity for referable DR.

Patel et al. (2023) [6] Addressed class imbalance using over/undersampling and class-weighted loss. A two-stage transfer learning approach with DenseNet169 offered the best balance of accuracy (91.7%) and computational efficiency.
Espinosa-Leal and Åberg (2023) [7] benchmarked preprocessing across 5 architectures. Most effective workflow: cropping to retinal ROI, CLAHE, and vessel enhancement improved accuracy by 3.2-8.7%.

III. Method

A. *Data Acquisition and Exploratory Data Analysis (EDA)*

The dataset, sourced from a Kaggle competition (Diabetic Retinopathy Classification #3), comprises 2,197 training and 1,465 test retinal fundus images. These images are categorized into five classes corresponding to different stages of diabetic retinopathy:

- **Class 0:** No DR
- **Class 1:** Mild DR
- **Class 2:** Moderate DR
- **Class 3:** Severe DR
- **Class 4:** Proliferative DR

Initial exploratory data analysis involved examining the distribution of images across these five classes as it can be seen in **Fig. 2** to understand the inherent class imbalance.

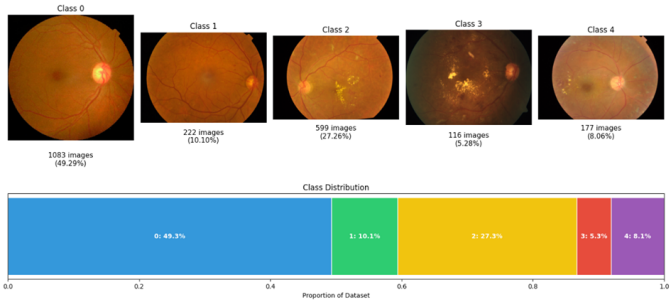

Fig. 2: Class Distribution of Retinal Fungus Images

B. *Image Preprocessing and Augmentation*

Given the varied nature of the fundus images and the common challenges in medical imaging (e.g., varying lighting, noise, subtle features), robust preprocessing and augmentation strategies were critical.

*1) **Exploration of Specific Image Enhancement Techniques**:* We qualitatively assessed several image enhancement techniques to highlight DR features, as illustrated in **Fig. 3**. These included:

- **Original Image:** Original image for visual comparison.
- **Edge Detection:** Emphasizes retinal edges, useful for outlining vessels and lesions visually.
- **Heatmap Visualization:** Highlights areas of intensity variation, indicating potential regions of interest or pathology through color mapping.

  CLAHE and green channel enhancement were implemented into the main preprocessing pipeline based on their benefit, while others primarily served for deeper understanding rather than direct model input.

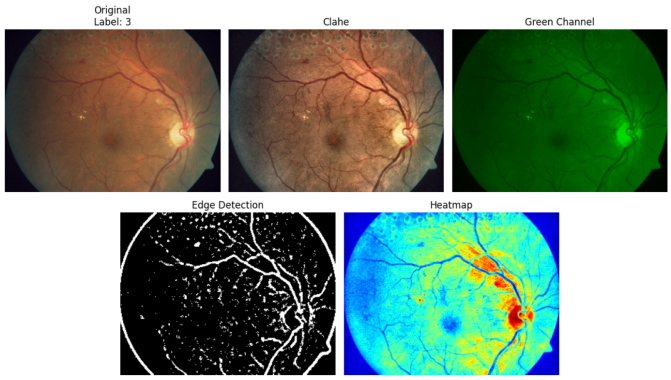

Fig. 3: Image Enhancement Techniques

*2) **Standard Preprocessing and Augmentation Pipeline** :*

- **Contrast Limited Adaptive Histogram Equalization (CLAHE):** CLAHE operates on the L-channel (Lightness) of the LAB color space, which effectively brightens darker regions and darkens brighter regions, making subtle features like microaneurysms and hemorrhages more visible. A clip limit of 2.0 (or 3.0/4.0 in some iterations) and a tile grid size of 8x8 were used. (Applied for Models 2, 3, 4, and 5).
- **Green Channel Enhancement:** Used for Model 1 (EfficientNetB3) to improve vessel and lesion contrast. It is done by extracting the green channel and replicating it across all three channels.
- **Resizing:** Images uniformly resized (224x224, 380x380, or 384x384 pixels) based on model input requirements.
- **Gaussian Blur with Sharpening (Ben Graham's Method):** . It sharpens the image by subtracting a blurred version from a scaled original image, enhancing fine details and edges of lesions. Applied for ViT-based models (Models 4 and 5).
- **Circular Crop:** Explored in some iterations to focus on the fundus region. Didn't improve the accuracy.
- **Normalization:** Pixel values were normalized to a range of [0,1] by dividing by 255.0. For PyTorch models, ImageNet-based normalization was applied: mean values of [0.485, 0.456, 0.406] and standard deviation values of [0.229, 0.224, 0.225].

To increase dataset diversity and reduce overfitting, various data augmentation techniques were applied during training. Geometric augmentations including horizontal and vertical flips, random rotations, and affine transformations.

C. *Model Architectures and Training Strategies*

We experimented with different deep learning architectures and multi-stage classification pipelines:

1) **Model 1: EfficientNetB3**: This model uses a pre-trained EfficientNetB3 for direct 5-class prediction, using Categorical Cross-Entropy loss. Input images undergo pre-

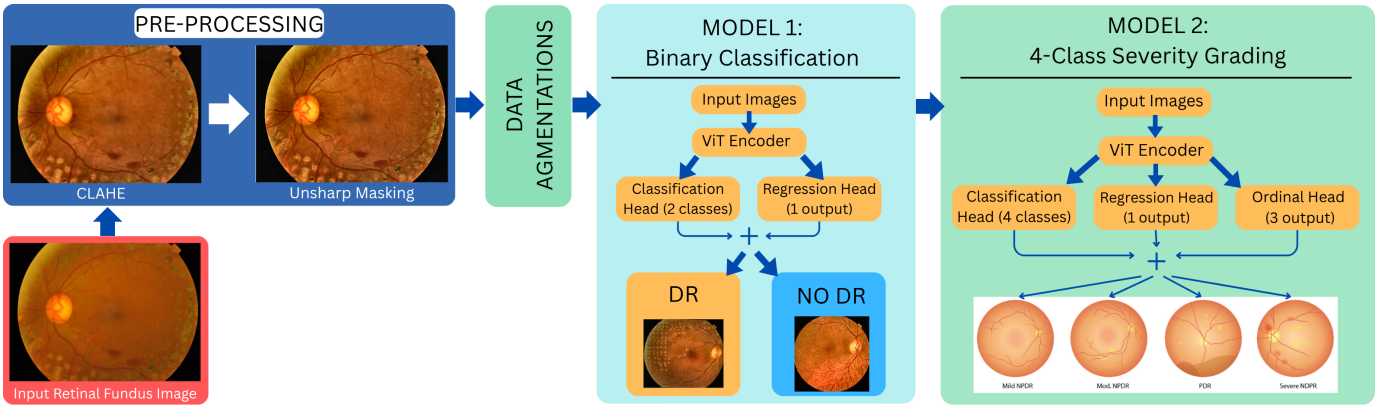

Fig. 4: Model 5 Pipeline

processing steps like green channel enhancement, circular cropping, resizing and normalization of pixel values. The model was trained over 30 epochs.

2) **Model 2: Three-Headed Method**: Inspired by multi-task learning, this model used a shared EfficientNet-B4 encoder feeding into three distinct heads (classification, regression, ordinal), with a separate Combiner Model integrating their outputs for final prediction. The utilized loss functions are: Weighted Cross Entropy, MSE, BCE-WithLogitsLoss. Several preprocessing steps are applied: CLAHE, circular cropping, resizing and normalization. It's trained for 75 epochs for each fold across 5-fold cross-validation. Additionally, at the end of each fold, the combined model of three heads is further trained for 5 epochs.

3) **Model 3: EfficientNet-B4 with Focal Loss**: A variation of Model 2, focusing solely on 5-class classification using an EfficientNet-B4 and Focal Loss to address class imbalance. Input images undergo the same preprocessing as Model 2. The model is trained for 75 epochs across 5-fold cross-validation.

4) **Model 4: Two-Stage ViT Model**: This cascaded approach used Vision Transformer (ViT) models in two stages:

   a) **Stage 1 (Binary Classification):** A MultiTaskModel (ViT backbone, classification, and regression heads) was trained on the full dataset to classify "No DR" vs. "DR" using Cross-Entropy and MSE loss. Trained for 10 epochs.

   b) **Stage 2 (4-Class Severity Grading):** A separate MultiTaskModel was trained only on images identified as having DR to classify 4 severity grades using Weighted Cross-Entropy and MSE loss. Trained for 25 epochs.

   Input images undergo these preprocessing steps: CLAHE, Ben Graham's Method, resizing and normalization.

5) **Model 5: Two-Stage ViT Model with Three-Headed Stage 2**: An evolution of Model 4 with same preprocessing steps, this model used the two-stage ViT pipeline but implemented the Three-Headed architecture (e.g. **Fig. 4**) for the second stage, and used Focal Loss for classification in both stages. The prediction process was identical to Model 4. this model was trained for 15 epochs on Stage 1 and 30 epochs on Stage 2.

Ensemble Methods: All models used K-Fold Cross-Validation, with test predictions obtained through:

- **Soft-Voting:** Averaging probabilities from individual folds (and TTA variants) for classification outputs.
- **Trimmed Mean (0.25):** Applied to Model 2's raw regression scores across folds to reduce outlier impact.

IV. Experimental Results

This section details the performance analysis of our deep learning solutions for diabetic retinopathy classification.

A. *Evaluation Metrics*

In addition to traditional accuracy which was only used for Model 1, we primarily evaluate model performance using Cohen's Kappa statistic for other four models. Kappa is a metric for assessing predictions-true labels agreement in classification. It is valuable when dealing with imbalanced datasets like the one in this project. Unlike simple accuracy, Kappa accounts for the agreement occurring by chance, providing a more reliable measure of a model's true predictive power.

The formula for Cohen's Kappa is given by:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

Where:

- $\mathbf{p_o}$ is the observed agreement (accuracy)
- $\mathbf{p_e}$ is the hypothetical probability of chance agreement

Kappa values typically range from -1 to 1, where:

- 1: indicates perfect agreement.
- 0: indicates agreement equivalent to chance.
- Negative values indicate agreement worse than chance.

For this project, higher Kappa values on the validation sets are indicative of better model performance in distinguishing between different DR severity levels.

B. Overall Performance and Discussion

Results revealed a clear progression in model performance:

Model 1 (EfficientNetB3): Achieved a Kaggle accuracy of **0.78020** with validation accuracy of 0.9922. Its relatively lower test performance compared to high training accuracy suggests some overfitting.

Model 2 (Three-Headed Method): Significantly improved Kaggle accuracy to **0.83412**. Models best validation Kappa scores ranged from 0.8948 to 0.9268.

Model 3 (EfficientNet-B4 with Focal Loss): Boosted Kaggle accuracy to **0.84163** by utilizing Focal Loss, which effectively handled class imbalance and improved classification for rarer severe DR classes. Best validation Kappa scores ranged from 0.8882 to 0.9167.

Model 4 (Two-Stage ViT Model): Achieved **0.85870**. Its Stage 1 (binary DR detection) consistently had high validation Kappa (0.9499 to 0.9773), with Stage 2 (4-class severity) further refining predictions (Kappa 0.5930 to 0.6817). This confirmed the effectiveness of problem decomposition.

Model 5 (Two-Stage ViT Model with Three-Headed Stage 2): Achieved the highest accuracy of **0.86075**. Stage 1 again showed strong validation Kappa (0.9590 to 0.9818), and Stage 2 Kappa (ranging from 0.5841 to 0.7438) generally improved over Model 4, validating the combined strategy. The training loss and validation Kappa of Stage 1 and 2 for Model 5, best performing model, can be seen in Fig. 5.

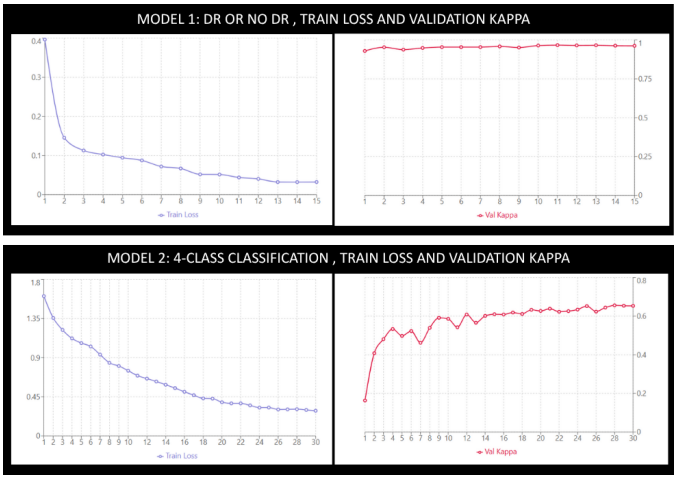

Fig. 5: Model 1 & 2 - Training Loss and Validation Kappa

C. Qualitative Analysis and Findings

The progressive improvements in accuracy across the models suggest several key findings:

1) **Addressing Class Imbalance:** The substantial gains observed with the introduction of Focal Loss (Model 3 and Model 5) underscore its importance in medical imaging where class distributions are often skewed. It allows the models to learn more effectively from underrepresented severe cases.
2) **Superiority of Transformer Architectures:** The clear performance leap when moving from EfficientNet to Vision Transformer (ViT) backbones (Models 4 and 5) indicates that these architectures are highly effective for capturing long-range dependencies and global contextual information in retinal images.
3) **Effectiveness of Two-Stage Classification:** The cascaded approach (Models 4 and 5) simplifies the overall problem. By first accurately detecting the presence of DR, the second stage can then focus on the more challenging task of severity grading among DR-positive cases. This decomposition of the problem appears to improve overall system performance.

V. Conclusions

This report explored deep learning methodologies for automated Diabetic Retinopathy (DR) severity classification from retinal fundus images, successfully addressing complexities like class imbalance.

Our five models demonstrated progressive improvements, highlighting the importance of robust preprocessing (e.g., CLAHE, Ben Graham's method), effective class imbalance handling (Focal Loss), and the benefits of multi-task learning for richer feature representations. A significant performance leap was observed with Vision Transformer (ViT) architectures and a two-stage classification approach, which simplified the problem and improved accuracy.

While our most advanced model (Model 5) achieved a high Kaggle accuracy of 0.86075, securing 2nd place in the competition, future work is needed to improve detection of subtle early DR signs, handle ambiguous cases, and ensure robust generalizability across diverse external datasets. Addressing these will be crucial for clinical translation.

References

[1] Alshahrani, M.; Al-Jabbar, M.; Senan, E.M.; Ahmed, I.A.; Saif, J.A.M. Hybrid Methods for Fundus Image Analysis for Diagnosis of Diabetic Retinopathy Development Stages Based on Fusion Features. Diagnostics 2023, 13, 2783.

[2] Che, Z., Yao, X., Wang, S., Ding, W., Gao, Z., Zhang, Y., & Yang, J. (2022). Application of deep learning in fundus image processing for ophthalmic disease diagnosis: A review. Frontiers in Endocrinology, 13, 1079217.

[3] Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2022). Diabetic retinopathy detection through deep learning techniques: A review. Informatics in Medicine Unlocked, 28, 100873.

[4] Pratiwi, M. I., Sari, D. K., & Wulandari, F. (2023). Classification of diabetic retinopathy using pre-trained CNN models with image augmentation. Diagnostics, 13(15), 2606.

[5] Fang, Y., Li, X., Hua, P., & Ma, F. (2024). Transformer-based efficient diagnosis approach for diabetic retinopathy screening. Annals of Medicine, 56(1), 2352018.

[6] Patel, Y., Shah, K., Panchal, D., & Prajapati, J. (2023). Detecting diabetic eye diseases using deep learning. International Journal for Research in Applied Science and Engineering Technology, 11(3), 1532-1538.

[7] Espinosa-Leal, L., & Åberg, F. (2023). Benchmarking deep learning models for classification of diabetic retinopathy. Applied Sciences, 13(5), 3108.