Fundus Unimodal and Late Fusion Multimodal Diabetic Retinopathy Grading

Sara El-Ateif[1][a] and Ali Idri[1,2][b]

[1]*Software Project Management Research Team, ENSIAS, Mohammed V University in Rabat, Morocco*

[2]*AlKhwarizmi College of Computing, Mohammed VI Polytechnic University, Marrakech-Rhamna, Benguerir, Morocco*

Keywords: Diabetic Retinopathy, Fundus, Multiclass, Late Fusion, Efficientnet, Swin Transformer, Hybrid-EfficientNetB0-SwinTF.

Abstract: Diabetic Retinopathy (DR) is an eye disease with complications, if left untreated grow, split into four grades: mild, moderate, severe, and proliferative. We propose to (1) compare and evaluate three different recently used deep learning models: EfficientNet-B5, Swin Transformer, and Hybrid-EfficientNetB0-SwinTF (HES) on the APTOS 2019 dataset's fundus and early fused (EF) weighted gaussian blur fundus. (2) Evaluate three fine-tuning and pre-processing schemes on the best model. And (3) choose the best model-scheme per modality and perform late fusion on them to get the final DR grade. Results show that our best method, late fusion HES model, results in F1-socre of 81.21%, accuracy of 81.83%, and AUC of 96.30%. We propose using late fusion HES model in population-wide diagnosis to assist doctors in Morocco to reduce DR burden.

1 INTRODUCTION

Diabetic Retinopathy (DR) an eye disease caused by diabetes, as of 2015, has affected 1 over 3 people according to the Moroccan League for the Fight against Diabetes. In 2018, the Moroccan Ministry of Health and Social Protection announced that more than 2 million citizens aged 18 and over are diabetic, with 50% unaware of the disease, and an estimation of 15,000 children affected. To detect such a disease, doctors perform routine eye screening using Slit-lamp – a tool that is rarely found in middle-income countries. Recent advances in technology brought a respectably accurate and affordable tool called digital fundus photography (Begum et al., 2021). DR affects a high number of citizens and it keeps growing as the number of diabetic patients grows, diagnosing it early helps avoid blindness and further complications (Teo et al., 2021). The solution resides in a wide spread diagnosis which is difficult to perform using restricted resources. This fact pushed researchers to investigate the use of artificial intelligence to help increase and speed up the diagnostic process (Sebastian et al., 2023). Most frequently used deep learning (DL) models for DR grading (i.e., classification into no DR, mild, moderate, severe, or proliferative DR) listed in (Sebastian et al., 2023), between 2017 and 2022, are: ResNet, VGG, InceptionV3, DenseNet, and EfficientNet based models; with EfficientNet being more present in 2022. Using the APTOS 2019 dataset (APTOS) (*APTOS 2019 Blindness Detection | Kaggle*, n.d.) of fundus images for DR grading, (Shahriar Maswood et al., 2020) train EfficientNet-B5 by optimizing the Quadratic Weighted Kappa (QWK) score and Mean Squared Error, and achieve an accuracy of 94.02%. (Karki & Kulkarni, 2021) train EfficientNet (B1 to B6) using different image resolutions, select the best of these variants (i.e., B1, B2, B3, B5), and use a weighted ensemble of these variants for the final prediction to achieve a QWK of 0.924377. (Canayaz, 2022) train EfficientNet and DenseNet models for feature extraction (results in 512 features). Then feed these extracted features to the wrapper methods to reduce the number of these features. Finally, the newly chosen features are fed to the SVM and Random Forest algorithms for classification. Using the Kaggle DR dataset, (Lazuardi et al., 2020) preprocess the fundus images using the Contrast Limited Adaptive Histogram Equalization (CLAHE)

[a] https://orcid.org/0000-0003-1475-8851

[b] https://orcid.org/0000-1111-2222-3333

El-Ateif, S. and Idri, A.
Fundus Unimodal and Late Fusion Multimodal Diabetic Retinopathy Grading.
DOI: 10.5220/0012048700003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 219-226
ISBN: 978-989-758-664-4; ISSN: 2184-285X

and center-cropping steps and combine them with progressive resizing training process using the EfficientNet B4 and B5 architectures; and obtain an accuracy of 83.88%. Recent work leverages the robustness of vision transformer (ViT) models for DR grading using different schemes (Gu et al., 2023; Sun et al., 2021; Yao et al., 2022): (Sun et al., 2021) propose a novel lesion-aware transformer (LAT) using an encoder-decoder structure to jointly grade and discover lesions. The novelty resides in using weakly supervised lesion localization via the transformer decoder, and lesion region importance and diversity to learn specific lesion filters using only image-level labels. On the Messidror-1 dataset, LAT scored 96.3% on area under the curve (AUC). (Yao et al., 2022) develop FunSwin, a Swin Transformer (Liu et al., 2021) based model, and reduce data imbalance by introducing data enhancements through rotation, mirroring, CutMix (Yun et al., 2019) and MixUp (Zhang et al., 2018). FunSwin scores an accuracy of 84.12% on the Messidor dataset. Gu et al. (Gu et al., 2023) propose a model composed of two blocks: feature extraction block (FFB) and a grading prediction block (GPB). FFB uses the classical ViT model to extract the features from the fundus images with its fine-grained attention. The GBP module, generates class-specific features using spatial attention. The results on the DDR dataset (Li et al., 2019) are an accuracy of 82.35%. In practice, to gain further clarity, DR and fundus eye diseases diagnosis require the use of other medical imaging modalities: fluorescein angiography (FA), and Multicolor imaging (MC). Few research works have been developed to leverage some of these modalities (Hervella et al., 2022; Song et al., 2021; Tseng et al., 2020): (Tseng et al., 2020) propose two fusion architectures similar to ophthalmologist's diagnostic process, late fusion and two-stage early fusion. Using the lesion and severity classification models, the late fusion approach combines these two models in parallel using postprocessing; while the two-stage early fusion combines them sequentially. These two approaches are trained on three Taiwanese hospitals datasets for DR severity grading, evaluated on the Messidor-2 dataset and scored 84.29% in accuracy. (Song et al., 2021) perform Multicolor DR detection (DR, no DR) using the multimodal information bottleneck network (MMIB-Net). The MMIB-Net uses the information bottleneck theory to compute the joint representation of different imaging modalities collected using the Multicolor imaging tool a confocal scanning laser ophthalmoscope (cSLO) that generates 16 images (8 for each eye) composed of Blue, Green, Infrared Reflectance and Combined-Pseudo color fundus images. The proposed model uses ResNet-50 as backbone and scored an accuracy of 94.30% on private collected data. (Hervella et al., 2022) perform DR grading using the fundus and FA modalities, by first pre-training on the unlabeled multimodal Isfahan dataset using their proposed Multimodal Image Encoding (MIE) approach. Then, using the fundus modality from a labeled dataset (IDRiD and Messidor), the pre-trained model is fine-tuned for DR grading. Finally, after pre-training and fine-tuning, the neural network uses the fundus modality to perform DR severity classification. On IDRiD, MIE scores an accuracy of 65.05%.

From the studies listed previously, a trend of pre-processing, use of large or private datasets, and training of different architectures in parallel (model ensembles or fusion) emerges. Meanwhile, for multimodal models and datasets they are scarce and in most cases the data is private; which could be related to the high cost needed for collecting such diversified modalities, storing them and ensuring patients data privacy. Another new emerging trend resides in the use of ViT models, due to their robustness and efficient attention mechanisms. But the classical ViT architectures require large amounts of data and computing resources to train; enters Swin Transformer, a ViT architecture optimized for relatively small data and less demanding computing resources. In this work we propose to leverage all of the previously used powerful elements (i.e., preprocessing, relatively small dataset, ViT, and multimodality) to train and compare the most recently used EfficientNet architecture with Swin Transformer and their hybrid combination: Hybrid-EfficentNetB0-SwinTF (Henkel, 2021); using the fundus, and the fusion of the Weighted Gaussian Blur Fundus (WGBF) (El-Ateif & Idri, 2022) with the fundus modality using the Discrete Wavelet Transformation (Naik & Kunchur, 2020) to obtain the early fused (EF) Fundus-WGBF modality. We evaluate the performance of these three models on this two different modalities using (1) six metrics: accuracy, sensitivity, specificity, F1-score, and AUC; along with (2) Scott-Knott Effect Size Difference (SK-ESD) (Tantithamthavorn et al., 2019) statistical test to cluster these models and find the best in terms of F1-score; on (3) the APTOS dataset. Then we further compare the best performing under different settings: fine-tuning, and preprocessing. Finally, we choose the most performant models per modality (fundus and EF), fine-tune them, and combine them using the late fusion approach.

In brief, the objective of this study is to evaluate whether the integration of preprocessing techniques with multimodality and image fusion, trained on a mid-sized dataset, could enhance the overall performance.

The remainder of this paper is organized as follows: Section 2 presents the data used and its preparation process. In section 3, we detail our approach. Section 4 lists our results and their discussion. Finally, section 5 concludes the paper and lays out our future work.

2 DATA PREPARATION

The models are trained and evaluated on the APTOS 2019 Blindness Detection dataset with the publicly available labels. The dataset contains fundus images provided by Aravind Eye Hospital in India and is split into 5 DR grades: (grade 0) No DR counts 1,805, (grade 1) Mild counts 370, (grade 2) Moderate counts 999, (grade 3) Severe counts 193, and (grade 4) Proliferative DR counts 295 fundus images. Note that for the models training we split these reported numbers into: train (80% of 2930: 2344), validation (20% of 2930: 586) and test (20%: 732) sets for the 1st and 2nd training phases, as for the 3rd phase we use only the train (80%: 2930) and test (20%: 732) sets; while replicating the class imbalance by using scikit-learn stratified k-fold method. Additionally, we resize the fundus images to 512x512 and remove the black background to leave the fundus only. Moreover, to avoid overfitting we perform for the Hybrid-EfficentNetB0-SwinTF and late fusion models the following data augmentations: horizontal flip, 0.1 height and width shift range, 20° rotation, 0.1 shear range, and 0.1 zoom range. Finally, for all the models, we normalize the data by dividing the pixels by 255 to keep them into the range of 0 to 255.

In the following we detail how we generated the second modality early fused (EF) Fundus-WGBF modality and expand on the pre-processing performed to experiment with performance enhancement of the best classified models for the fundus and EF modalities. Figure 1 showcases all of these processes by DR grade: fundus, its pre-processing, WGBF, EF and its pre-processing.

2.1 Second Modality

To generate the early fused (EF) Fundus-WGBF modality we at first generate the Weighted Gaussian Blur Fundus (WGBF) from the fundus images using the method followed in (El-Ateif & Idri, 2022): (1) crop the image to remove the black background, (2) resize the image to 512x512, and (3) apply the Graham algorithm (Graham, 2015) that serves to subtract the local average color and map 50% gray to this local average so as to enhance the blood vessels and DR lesions. After that we apply the Discrete Wavelet Transformation (DWT) algorithm (Naik & Kunchur, 2020) with the mean method using as input the fundus and WGBF images of a respective patient to obtain the early fused (EF) Fundus-WGBF modality. The DWT helps capture the best aspect from the fundus (color distribution and fundus structure) and WGBF modalities (i.e., enhanced blood vessels and DR lesions), respectively.

2.2 Pre-Processing

In previous work (Canayaz, 2022; Karki & Kulkarni, 2021; Lazuardi et al., 2020; Shahriar Maswood et al., 2020), pre-processing the fundus images helped improve significantly the studied models performance by enhancing the images quality. To test if the models trained replicate the same results, we perform the following: For the fundus modality we apply: (1) median blur (aperture=5), (2) gamma correction (gamma=1.7), and (3) CLAHE with clip limit equal to 2.0 and tile gride size of 80 by 80. For the EF modality, we apply: (1) gamma correction (gamma=1.5), and (2) CLAHE with clip limit equal to 2.0 and tile gride size of 60 by 60.

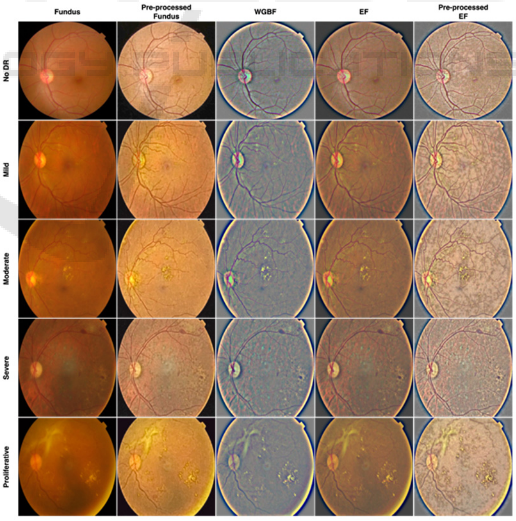

Figure 1: Data samples of fundus, pre-processed fundus, WGBF, EF, and pre-processed EF by DR grade.

3 EXPERIMENTAL SETUP

We train and evaluate three different models: EfficientNet-B5 (ENetB5), Swin Transformer (STF), and Hybrid-EfficientNetB0-SwinTF (HES); using the fundus, and EF modalities. We perform this training in three phases: Phase 1: We train using 5-fold stratified cross validation (CV) on the train, and validation sets EfficientNet-B5, Swin Transformer, and Hybrid-EfficientNetB0-SwinTF for each of the fundus and EF, respectively. Then we evaluate and compare on the 5-fold CV test set; and pick the two best models, based on the F1-score, figuring in 1st and 2nd cluster of SK-ESD. Phase II: The best models per modality (i.e., Fundus HES, and EF HES) are then compared and trained on the 5th fold based on accuracy following four schemes for each modality: (1) fine-tuning using weights saved in Phase I (FT0), (2) fine-tuning using weights from Phase I while unfreezing the last 20 layers of ENetB0 (FT1), (3) fine-tuning using weights from Phase I while unfreezing the last 20 layers of ENetB0 and freezing STF blocks (FT2), (4) training on pre-processed fundus and EF data. Phase III: We take the best from these four schemes (i.e., FT2 Fundus (no pre-processing train and test) and EF (pre-processed train and test) HES model) and use the FT2 scheme to train these two models in parallel and average their predictions to get the final DR grade, to perform what is called late or decision level fusion. Hereafter, we detail the four models' architectures with their training and testing processes, the empirical process, statistical methods and performance metrics used to evaluate the models.

3.1 Modeling

We use the ImageNet pre-trained EfficientNet-B5 (Tan & Le, 2019) (ENetB5) from the EfficientNet family. EfficientNet models are convolutional neural networks with uniform scaling faculties of width, depth, resolution using compound coefficients. In the classification layer of this model, we use: Global Average Pooling 2D, Dropout (0.5), Dense (1024 nodes, and ReLU as activation), and the output layer composed of a Dense layer with 5 nodes and Softmax as activation function. We freeze all of its layers apart from the classification part and train using Adam optimizer and categorical cross entropy as loss. ENetB5 uses as input 456x456 images and batch size of 32.

From the ViT family of models we use Swin Transformer (Liu et al., 2021) (STF) a hierarchical ViT using shifted windows scheme to improve self-attention computation efficiency to non-overlapping local windows and provide cross-window connection. The model takes 224x224 in input shape, 4x4 in patch size, 0.03 in dropout rate, 8 in number of heads, 96 in embedding dimension, 1024 in number of nodes for multilayer perceptron (MLP), 7 in window size, 1 in shift size, 0.1 in label smoothing for the categorical cross entropy loss, and batch size of 32.

The 3rd model we train is Hybrid-EfficientNetB0-SwinTF (HES) (Henkel, 2021), see Figure 2, composed of ImageNet pre-trained ENetB0 using the AdvProp (Xie et al., 2020)–an adversarial training approach used to prevent overfitting by treating adversarial examples as additional examples–weights in its head, STF in its body, and a Dense layer with 5 nodes and Softmax activation as output. In the classification head of STF we define: Global Average Pooling 1D, Alpha Dropout (0.5), and Batch Normalization. While for ENetB0 we define: Global Average Pooling 2D, Alpha Dropout (0.5) and use sparse categorical cross entropy as loss with no label smoothing. The model takes as input 384x384 sized images, batch size of 8, patch size of 2x2, dropout of 0.5, 8 in number of heads, embedding dimension of 64, 128 for MLP, 2 in window size, 1 in shift size and 24x24 STF image dimension. We retrieve 'block6a_expand_activation' output of the ENetB0 model to reduce the image size of fundus and EF from 384x384 to 24x24 to reduce features size and get a better representation of DR grades. Both HES and STF models use AdamW optimizer with 1e-3 in learning rate, and 0.0001 in weight decay.

For the late fusion model, we use FT2 of HES of fundus and EF, train them on train and test split (2930 train and 732 test) in parallel and average their predictions to get the final output. But for EF HES we use the SGD optimizer (learning rate of 1e-4) instead of AdamW and train on pre-processed EF with their respective weights.

All models are trained for 50 epochs, using reduce learning rate on plateau and early stopping that monitor the validation loss.

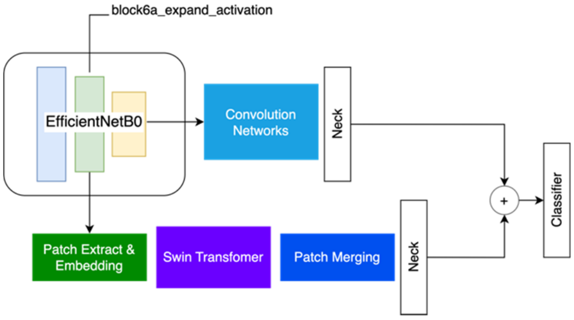

Figure 2: Hybrid-EfficientNetB0-SwinTF (HES) model architecture.

3.2 Empirical Process

The empirical process followed is split into three phases:

1) Phase I: Cluster the fundus and EF three models (ENetB5, STF, HES) using SK-ESD based on F1-score to select the best model by modality.
2) Phase II: Train and compare the best models from (1) following FT0, FT1, FT2, pre-processing schemes using accuracy as performance metric.
3) Phase III: Pick the best approach from (2) for fundus and EF respectively, train them in parallel, average their predictions and report sensitivity, specificity, precision, and F1-score by DR grade and accuracy and weighted average of one versus rest AUC.

3.3 Statistical Methods and Performance Metrics

We hope you find the information in this template useful in the preparation of your submission. In phase I we trained and evaluated the three models using 5-fold stratified CV, and clustered the models by modality using Scott-Knott Effect Size Difference (SK-ESD) (Tantithamthavorn et al., 2019) statistical test based on F1-score. In phase II we reported the accuracy values of the 5th fold. In phase III, we reported the scores of six metrics for each DR grading to evaluate the performance of the best models: accuracy, sensitivity, specificity, precision, F1-score, and weighted average one versus rest AUC score. These first five metrics are defined in Eqs.1–5 respectively:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TN+TP+FP+FN}} \quad (1)$$

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{F1} = 2\ \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

where: TP: true positive. FP: false positive. TN: true negative, and FN: false negative.

4 RESULTS AND DISCUSSION

In this section we lay out the results of phase I, II, and III for the three models ENetB5, STF and HES (phase I), HES fundus and EF trained using different fine-tuning schemes and pre-processing (phase II), as well as late fusion HES of best fundus and EF HES (phase III). Figure 3 lays out the results of SK-ESD clustering of all three trained models from phase I by modality based on F1-score (to account for class imbalance). Table 1 reports accuracy scores and loss of phase II experiments. Table 2, 3 and 4 record the results of best results from phase II (Table 2 and 3) of FT2 HES and phase III (Table 4) late fusion HES using precision, sensitivity, F1-score, and specificity by DR grade. Table 5 lays out the comparison between our best proposed model and state-of-the-art trained on the APTOS dataset. In the following we layout these results and discuss what they entail.

4.1 Unimodal Results

Figure 3 shows that the best ranking model according to SK-ESD is Fund_HES or fundus Hybrid-EfficentNetB0-SwinTF as it appears in the first cluster. In second place we find the EF_HES or EF Hybrid-EfficentNetB0-SwinTF model. Apart from HES and ENetB5 models where the fundus outperformed EF, for STF the EF modality outperformed the fundus one. This result may be related to the propriety of ViT models that are robust to changes in images. In conclusion, for phase I, the best ranking models are fundus HES and EF HES with an F1-score of 78.62% and 76.53%, respectively.

For phase II, we train fundus and EF HES models using different schemes of fine-tuning and pre-processing of fundus and EF by using saved weights from phase I HES per modality. From Table 1, the best approach reported in terms of accuracy and loss is HES model trained using pre-processed EF with Stochastic Gradient Descent (SGD) optimizer instead of AdamW (Loshchilov & Hutter, 2019) and non-pre-processed fundus with previously set AdamW optimizer (as we previously trained it after pre-processing and by changing to SGD optimizer and it provided worse results) on the train set and evaluated on the validation and test sets with an accuracy of 92.35% for fundus and 88.39% for EF. The change of the optimizer helped improve the training process by improving the research for optimal parameters by slowly decreasing the learning rate through SGD. As for EF pre-processing, it improved the DR lesions detection by smoothing the features and equilibrating the color distribution. Meanwhile, FT2 approach helped train the ENetB0 module further for better feature extraction using convolutions while keeping the best representation from Swin blocks.

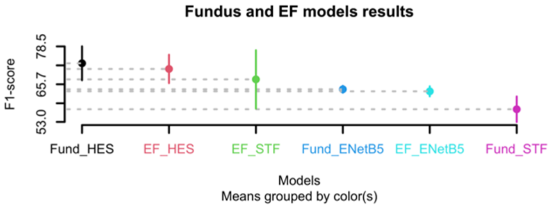

Figure 3: Fundus and EF SK-ESD F1-score results, with Fund referring to the fundus modality and EF referring to the early fused Fundus-WGBF DWT modality.

Table 1: Accuracy in % (loss) results for HES model on fundus and EF modalities, respectively.

Phase	Fundus HES (%)	EF HES (%)
II-Pre-processing on Train: Val: Test	72.40 (2.06)	73.50 (2.16)
II-FT0 on Train: Val: Test	73.50 (2.55)	86.75 (1.14)
II-FT1 on Train: Val: Test	89.10 (1.05)	87.84 (0.93)
II-FT2 on Train: Val: Test	92.35 (0.46)	86.75 (0.87)
II-FT2 (PrepEF with SGD*) on Train: Val: Test	92.35 (0.46)	88.39 (0.69)
I on Train: Test	76.23 (1.88)	75.82 (1.75)
II-FT2 (PrepEF with SGD*) on Train: Test	81.83 (1.53)	79.09 (0.77)

*PrepEF stands for pre-processed EF and SGD references to using SGD optimizer.

4.2 Late Fusion Results

In phase III, we retrained the fundus HES (weighted F1-score = 76%) and EF HES (weighted F1-score = 77%) on the train and test sets (the validation set was added to the train set) on the 5th fold so as to get the respective weights to train the late fusion model. The late fusion HES model consists of training in parallel HES fundus (with AdamW) and HES pre-processed EF (with SGD) using these new weights and following the FT2 scheme. Finally recuperating the predictions from FT2-HES-Fundus and FT2-HES-PrepEF and averaging them to get the final DR grading. As exposed in Table 2, 3 and 4, the F1-score of late fusion HES equal to 81.21% (accuracy=81.83%, AUC=96.30%) is better than the average F1-score of phase I (on Train: Test) fundus and EF HES equal to 76.50% (accuracy=76.03%, AUC=94.55%) and average F1-score of phase II-FT2 fundus and PrepEF equal to 60.96% (accuracy = 80,46%, AUC=89,81%). Table 2 and 3 show that each model II-FT2 HES of fundus and EF modalities compensated for the weaknesses of the other model through the late fusion approach. Although, the scores for the mild, severe, and proliferative are still below 70%; but based on the number of samples available for these classes (74/732, 39/732, 59/732, respectively) the provided results are promising.

Table 2: Metrics results of FT2 fundus HES model per DR grade.

DR grade	Precision (%)	Specificity (%)	F1-score (%)	Sensitivity (%)
No DR	97.52	97.57	97.79	98.06
Mild	43.09	89.36	53.81	71.62
Moderate	77.14	94.00	63.72	54.27
Severe	27.78	92.50	36.04	51.28
Proliferative	67.65	98.37	49.46	38.98
Weighted average	80.35	95.56	76.90	76.23

Table 3: Metrics results of FT2 pre-processed EF HES model per DR grade.

DR grade	Precision (%)	Specificity (%)	F1-score (%)	Sensitivity (%)
No DR	64.45	46.63	78.21	99.45
Mild	17.21	84.65	21.43	28.39
Moderate	100.00	100.00	04.90	02.51
Severe	25.00	96.97	20.90	17.95
Proliferative	45.00	98.37	22.78	15.25
Weighted average	65.67	71.83	45.02	54.78

Table 4: Metrics results of late fusion HES model per DR grade.

DR grade	Precision (%)	Specificity (%)	F1-score (%)	Sensitivity (%)
No DR	96.75	96.77	97.80	98.90
Mild	51.37	91.94	61.20	75.68
Moderate	77.35	92.31	73.68	70.35
Severe	52.94	98.85	32.14	23.10
Proliferative	64.29	97.03	62.61	61.02
Weighted average	81.94	95.20	81.21	81.70

4.3 Comparison with State-of-the-Art

In terms of accuracy compared to state-of-the-art model (Canayaz, 2022) that used nature-inspired wrappers with EfficientNet and scored an accuracy of 96.32%, our best model late fusion HES accuracy is less than ~14.49% on the APTOS dataset. The difference in performance could be due to the interesting method presented by (Canayaz, 2022) along with the different pre-processing approach undertaken (i.e., masks created according to a tolerance value). As for the second-best model (Shahriar Maswood et al., 2020), that train to optimize the QWK score and Mean Squared Error, it outperformed our model by 11.50%. From these results we conclude that leveraging preprocessing, multimodality, and image fusion does not result into model performance improvement as would be expected. But, as transformer models are more robust in comparison with convolution models, further development and validation on real world datasets is still needed for further conclusions.

Table 5: State-of-the art comparison with best resulting model using the APTOS dataset.

Model	Modality	Accuracy (%)
(Canayaz, 2022)	Fundus	96.32
(Shahriar Maswood et al., 2020)	Fundus	93.33
Late Fusion HES (proposed)	PrepEF	81.83

5 CONCLUSION AND FUTURE WORK

In this work we proposed to train and compare three different models: EfficientNetB5, Swin Transformer, and Hybrid-EfficientNetB0-SwinTF according to three phases and using different schemes (fine-tuning and pre-processing) for the fundus and early fused fundus with its preprocessing. The best models from fundus and early fused modality were trained in parallel and their predictions were averaged to predict the final DR grade as a late fusion process. Although the results were promising with an accuracy of 81.83%, compared to state-of-the-art (Canayaz, 2022) (accuracy=96.32%), our late fusion model still needs fine-tuning. In future work, we aim to improve: the data quality by introducing a more diverse and grades enriched dataset from different hospitals, the pre-processing process by proposing a dynamic approach (changes depending on image quality), and optimization of the late fusion approach by reducing the numbers of parameters; and perform further validations on local datasets.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude for the support provided by Google Ph.D. Fellowship.

REFERENCES

APTOS 2019 Blindness Detection | Kaggle. (n.d.). Retrieved November 1, 2021, from https://www.kaggle.com/c/aptos2019-blindness-detection/overview/description

Begum, T., Rahman, A., Nomani, D., Mamun, A., Adams, A., Islam, S., Khair, Z., Khair, Z., & Anwar, I. (2021). Diagnostic accuracy of detecting diabetic retinopathy by using digital fundus photographs in the peripheral health facilities of Bangladesh: Validation study. *JMIR Public Health and Surveillance*, *7*(3). https://doi.org/10.2196/23538

Canayaz, M. (2022). Classification of diabetic retinopathy with feature selection over deep features using nature-inspired wrapper methods. *Applied Soft Computing*, *128*, 109462. https://doi.org/10.1016/j.asoc.2022.109462

El-Ateif, S., & Idri, A. (2022). Single-modality and joint fusion deep learning for diabetic retinopathy diagnosis. *Scientific African*, *17*, e01280. https://doi.org/10.1016/j.sciaf.2022.e01280

Graham, B. (2015). *Kaggle Diabetic Retinopathy Detection competition report*. 1–9.

Gu, Z., Li, Y., Wang, Z., Kan, J., Shu, J., & Wang, Q. (2023). Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention. *Computational Intelligence and Neuroscience*, *2023*, 1–12. https://doi.org/10.1155/2023/1305583

Henkel, C. (2021). *Efficient large-scale image retrieval with deep feature orthogonality and Hybrid-Swin-Transformers*. 1–5. http://arxiv.org/abs/2110.03786

Hervella, Á. S., Rouco, J., Novo, J., & Ortega, M. (2022). Multimodal image encoding pre-training for diabetic retinopathy grading. *Computers in Biology and Medicine*, *143*. https://doi.org/10.1016/j.compbiomed.2022.105302

Karki, S. S., & Kulkarni, P. (2021). Diabetic retinopathy classification using a combination of EfficientNets. *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, *Figure 2*, 68–72. https://doi.org/10.1109/ESCI50559.2021.9397035

Lazuardi, R. N., Abiwinanda, N., Suryawan, T. H., Hanif, M., & Handayani, A. (2020). Automatic diabetic retinopathy classification with efficientnet. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, *2020-Novem*, 756–760. https://doi.org/10.1109/TENCON50793.2020.9293941

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., & Kang, H. (2019). Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, *501*, 511–522. https://doi.org/10.1016/j.ins.2019.06.011

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE International Conference on Computer Vision*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*.

Naik, R. B., & Kunchur, P. N. (2020). Image Fusion Based on Wavelet Transformation. *International Journal of Engineering and Advanced Technology*, *9*(5), 473–477. https://doi.org/10.35940/ijeat.D9161.069520

Sebastian, A., Elharrouss, O., Al-Maadeed, S., & Almaadeed, N. (2023). A Survey on Deep-Learning-Based Diabetic Retinopathy Classification. *Diagnostics*, *13*(3), 345. https://doi.org/10.3390/diagnostics13030345

Shahriar Maswood, M. M., Hussain, T., Khan, M. B., Islam, M. T., & Alharbi, A. G. (2020). CNN Based Detection of the Severity of Diabetic Retinopathy from the Fundus Photography using EfficientNet-B5. *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, IEMCON 2020*, 147–150. https://doi.org/10.1109/IEMCON51383.2020.9284944

Song, J., Zheng, Y., Wang, J., Zakir Ullah, M., & Jiao, W. (2021). Multicolor image classification using the multimodal information bottleneck network (MMIB-Net) for detecting diabetic retinopathy. *Optics Express*, *29*(14), 22732. https://doi.org/10.1364/oe.430508

Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., & Zhang, Y. (2021). Lesion-Aware Transformers for Diabetic Retinopathy Grading. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10933–10942. https://doi.org/10.1109/CVPR46437.2021.01079

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 10691–10700. http://arxiv.org/abs/1905.11946

Tantithamthavorn, C., McIntosh, S., Hassan, A. E., & Matsumoto, K. (2019). The Impact of Automated Parameter Optimization on Defect Prediction Models. *IEEE Transactions on Software Engineering*, *45*(7), 683–711. https://doi.org/10.1109/TSE.2018.2794977

Teo, Z. L., Tham, Y.-C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., Wong, I. Y., Ting, D. S. W., Tan, G. S. W., Jonas, J. B., Sabanayagam, C., Wong, T. Y., & Cheng, C.-Y. (2021). Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045. *Ophthalmology*, *128*(11), 1580–1591. https://doi.org/10.1016/j.ophtha.2021.04.027

Tseng, V. S., Chen, C.-L., Liang, C.-M., Tai, M.-C., Liu, J.-T., Wu, P.-Y., Deng, M.-S., Lee, Y.-W., Huang, T.-Y., & Chen, Y.-H. (2020). Leveraging Multimodal Deep Learning Architecture with Retina Lesion Information to Detect Diabetic Retinopathy. *Translational Vision Science & Technology*, *9*(2), 41. https://doi.org/10.1167/tvst.9.2.41

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., & Le, Q. V. (2020). Adversarial examples improve image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 816–825. https://doi.org/10.1109/CVPR42600.2020.00090

Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., & Wang, Z. (2022). FunSwin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. *Frontiers in Physiology*, *13*(July), 1–9. https://doi.org/10.3389/fphys.2022.961386

Yun, S., Han, D., Chun, S., Oh, S. J., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 6022–6031. https://doi.org/10.1109/ICCV.2019.00612

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). MixUp: Beyond empirical risk minimization. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 1–13.