Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

An Ensemble Approach for Robust Diabetic Retinopathy Detection

Aman Panjwani, K. Nishanth Reddy, Ch. Gopi Chandu, E. Sai Kiran,
B.Tech Students of CSE(AI&ML)

Bhaskar Das
Associate Professor, Department of CSE (Data Science), Hyderabad
Institute of Technology and Management, Hyderabad, India

***Abstract*—Diabetic Retinopathy (DR) is a top cause of blind- ness and visual impairment around the world, and is especially common among patients with diabetes. Early detection ensures effective treatment and prevention of disease progression. In this paper, we present a deep learning-based system designed to automatically detect and grade DR using retinal fundus images. The system uses an ensemble of EfficientNet-based convolutional neural networks, trained in an ensemble framework using a synthetically created balanced dataset of retinal fundus images in the APTOS 2019 Blindness Detection dataset to counteract the impact of class imbalance. An extensive preprocessing scheme, comprising contrast enhancement, resize, and normalization, was used to optimize image quality and model resilience. Experimental results confirm that ensemble-based models achieve higher accuracy, sensitivity, and specificity at every level of DR severity when compared to multiple baseline schemes. The results underscore the possibility of using EfficientNet-based ensemble models as scalable and effective solutions for DR diagnosis, especially in resource-poor clinical environments.**

***Index Terms*—Ensemble Learning, EfficientNet, Convolutional Neural Networks, Fundus Imaging, APTOS 2019, Image Classification, Data Augmentation, Medical Imaging.**

I. INTRODUCTION

Diabetic Retinopathy (DR) is the most serious complica- tion of diabetes, occurring in more than 400 million people globally and still accounting for preventable blindness. The development of DR is associated with retinal vascular damage and clinically can be classified as no DR, mild, moderate, severe, and proliferative. Early detection and proper classification are crucial to avoid worsening of vision and enable proper medical treatment at an appropriate time. The conventional diagnosis is, however, heavily dependent on skilled ophthalmologists and, therefore, is time-consuming, subjective, and susceptible to inter-observer variation, especially in resource-poor environments.

Automated DR detection from retinal fundus images is now much more feasible thanks to recent developments in deep learning. In image-based medical diagnostics, deep convolutional neural networks (CNNs) have shown impressive capabilities, potentially leading to quicker and more reliable screening results. However, issues like class imbalance, inconsistent image quality, and minor inter-class variations still affect model performance.

A. Dataset Enhancement

To address data-related limitations, this study utilizes a synthetically balanced version of the APTOS 2019 Blindness Detection dataset. The dataset is enhanced through a targeted set of preprocessing and augmentation techniques designed to simulate real-world clinical variability. These include geometric transformations (e.g., random cropping, flipping, and rotation), artifact removal (circle cropping and black border removal), and image smoothing (Gaussian blur). All images are resized to a uniform dimension of **512×512** pixels and normalized to standardize the intensity distribution. This comprehensive preprocessing pipeline helps mitigate class imbalance, reduce overfitting, and improve the model's robustness to illumination and noise variations.

B. Ensemble Learning Strategy

To further improve classification performance, an ensemble-based deep learning approach is adopted. Specifically, the proposed framework integrates two EfficientNet variants—EfficientNet-B3 and EfficientNet-B4—each fine-tuned independently on the preprocessed dataset. These models serve as parallel base learners, capturing complementary representations of the input fundus images. A stacking-based ensemble method is employed, where the softmax probability vectors produced by each backbone are concatenated and passed through a meta-classifier implemented as a fully connected dense layer. This strategy enables the ensemble to learn an optimal combination of predictions, enhancing generalization and improving robustness across diverse imaging conditions.

Experimental results demonstrate that the proposed ensemble framework achieves superior performance in multi-class DR classification compared to individual base models, underscoring its potential for deployment in large-scale DR screening, especially in under-resourced healthcare environments.

II. RELATED WORK

A. Single-View DR Detection Methods

Early work in DR detection focused on CNN-based models trained on individual fundus images. Gulshan et al. pioneered this by leveraging InceptionV3 to classify over 120,000 retinal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

images with high diagnostic performance. Subsequent models improved feature extraction using deeper architectures like ResNet and EfficientNet. More recently, transformer-based methods, such as Vision Transformers (ViT) and FunSwin, have been employed to capture long-range dependencies and multi-scale representations in retinal images. However, these models are limited by their reliance on single-view images, making them prone to missing peripheral lesions and generating inconsistent predictions for ambiguous cases.

B. Multi-View DR Detection Methods

To overcome the limitations of single-view analysis, multi-view approaches have emerged. Takahashi et al. utilized GoogLeNet with multiple fundus angles to classify DR grades. Luo et al. extended this by introducing MVDRNet and later MVICINN, employing dual-branch CNN-transformer networks to extract both local and global features. These methods showed significant performance improvements but still suffered from redundant view information and poor feature fusion. To address this, newer methods like WGLIN introduced wavelet-based global-local interaction and cross-view attention modules that enhance lesion boundary learning and minimize feature redundancy.

C. Uncertainty-Aware Learning

While ensemble models have demonstrated higher accuracy and robustness, they often produce overconfident predictions. To mitigate this, uncertainty-aware models such as UATTA-ENS incorporate test-time augmentation and uncertainty-weighted ensemble learning. This allows the model to flag ambiguous cases, thereby reducing the risk of misdiagnosis. Metrics like Expected Calibration Error (ECE) and Brier Score are used to assess prediction reliability. These uncertainty-aware techniques provide an additional safety layer, particularly valuable in clinical applications.

III. METHODOLOGY

In this study, we propose an ensemble deep learning architecture for Diabetic Retinopathy (DR) detection, leveraging a stacking-based mechanism that integrates two complementary EfficientNet models—EfficientNet-B3 and EfficientNet-B4. Each model is fine-tuned independently on a synthetically augmented dataset to enhance feature diversity and generalization capability. The predictions from these models are subsequently combined using a meta-classifier, aiming to improve overall diagnostic accuracy and robustness across varying DR severity levels.

A. Dataset and Preprocessing

The APTOS 2019 Balanced Dataset is a curated version of the original APTOS 2019 Blindness Detection dataset that was used in this study. 9025 retinal fundus images make up this balanced dataset, and each one is classified into one of five groups that represent the stages of diabetic retinopathy (DR): no DR, mild DR, moderate DR, severe DR, and proliferative DR. The dataset was expanded so that the number of images in each class matched the maximum number of samples found across all categories in order to achieve class balance. The underrepresented classes' sample sizes were artificially increased using a variety of augmentation techniques, which produced a more consistent distribution throughout all DR stages. This method guarantees that the models developed using this dataset are not skewed toward any specific class and can learn robust features across all severity levels.

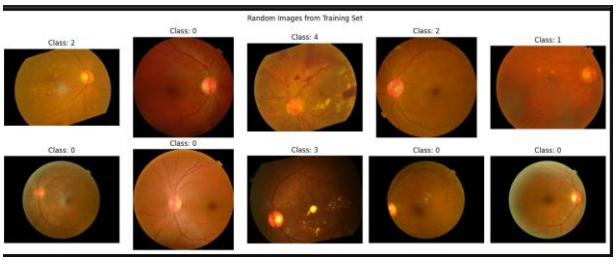

Fig. 1. Raw Images

To improve our ensemble model's robustness and generalization, we employed a comprehensive data augmentation pipeline. Given the variability in fundus image acquisition across devices, patients, and clinical settings, data augmentation plays a crucial role in simulating real-world imaging conditions and mitigating overfitting. The following augmentation techniques were applied during training:

- Random Horizontal and Vertical Flipping: Fundus im- ages may be captured with slight variations in orientation due to patient positioning or device handling. Flipping the images both horizontally and vertically ensures the model remains invariant to orientation shifts and learns symmetrical features more robustly.
- Rotation (0°–30°): Mild angular misalignments are common in fundus photography. Rotating the images ran- domly within a 30-degree range helps the model gen- eralize to such variations without degrading anatomical interpretability.
- Zoom and Scaling Transformations: Different fundus cameras and clinical protocols may result in images with varying zoom levels. By randomly zooming in and out, the model is trained to detect DR features across multiple spatial scales, thus enhancing its ability to localize both subtle microaneurysms and widespread hemorrhages.
- Random Cropping: This technique randomly selects sub-regions within the full retinal field, allowing the model to learn from localized patterns. This is particularly important in DR detection, where early-stage lesions may be small and confined to discrete regions.
- Circle Cropping: To focus the model's attention on the medically relevant central retinal area, circular cropping was employed. This removes irrelevant peripheral regions such as eyelid margins and vignetting artifacts, ensuring the network concentrates on diagnostically significant content.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

- Black Border Cropping: Most fundus images include black background margins that do not contribute to di- agnosis. These were cropped out to standardize image framing and improve the effective resolution of the retinal content presented to the model.
- Gaussian Noise and Gaussian Blur: Gaussian noise was introduced early in the augmentation process to improve the model's tolerance to low-quality imaging artifacts. In contrast, Gaussian blur was applied in later stages to simulate defocus or low-contrast conditions, enhancing the model's resilience to real-world degradation in image clarity.

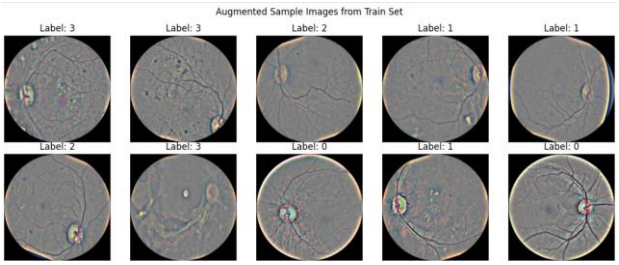

Fig. 2. Preprocessed Images

To maintain consistency throughout the dataset, all images were resized to 512 x 512 pixels after augmentation. Additionally, pixel intensity normalization was performed to stabilize the data distribution, leading to improved convergence during model training.

B. Model Architecture

1) Proposed Ensemble Architecture: The proposed model architecture is an ensemble-based deep learning system designed for the multi-class classification of Diabetic Retinopathy (DR) from retinal fundus images. It leverages two EfficientNet variants—EfficientNet-B3 and EfficientNet-B4—as parallel feature extractors. Both networks are initialized with ImageNet pre-trained weights, allowing the model to benefit from transfer learning, and are subsequently fine-tuned on a domain-specific DR dataset to better capture the unique characteristics of retinal images.

To adapt the base models for the five-class DR classification task—ranging from no DR to proliferative DR—the original classification heads of EfficientNet-B3 and EfficientNet-B4 are replaced with custom fully connected (FC) layers.To produce probabilistic class outputs, EfficientNet-B3 specifically connects its final 1536-dimensional feature vector to a dense layer that has five output neurons and a sigmoid activation function. Similarly, a corresponding FC layer and sigmoid activation are used to map the 1792-dimensional feature output of EfficientNet-B4 to a five-class probability distribution.

During inference, a given retinal fundus image is processed in parallel by both EfficientNet-B3 and EfficientNet-B4, each independently generating a five-dimensional probability vector corresponding to the five DR severity levels. These two vectors are concatenated to form a unified 10-dimensional representation, which serves as the input to a meta-classifier.

2) Meta-Classifier and Its Significance: The meta-classifier constitutes a critical component of the proposed ensemble framework. Rather than relying on heuristic combination strategies such as averaging or majority voting, the meta-classifier is implemented as a learnable function that fuses the predictions from both EfficientNet backbones in a data-driven manner. By training on the concatenated probability vectors, the meta-classifier is able to capture non-linear relationships and interdependencies between the outputs of EfficientNet-B3 and B4.

This fusion strategy allows the model to exploit complementary features learned by each backbone, thereby enhancing the system's overall discriminative power. The meta-classifier learns to down-weight conflicting signals and emphasize consistent predictive cues, thereby mitigating the biases and limitations inherent in each individual model. This method is particularly beneficial in the field of medical imaging, where trustworthy classification is severely hampered by intra-class similarity and inter-class variability.

Empirical evaluations demonstrate that this ensemble approach, bolstered by the meta-classifier, significantly outperforms single-model baselines in terms of accuracy and robustness. The architecture's ability to aggregate diverse feature representations contributes to improved diagnostic reliability, making it a promising solution for automated DR screening.

Fig. 3. Model Architecture

C. Training Procedure

The training methodology for the proposed deep learning framework involves a systematic, two-stage process to ensure effective learning and robust generalization. This section details the configurations, strategies, and rationale underlying the training pipeline.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

Initially, each backbone model—based on EfficientNet variants—is trained independently to perform the five-class classification of Diabetic Retinopathy (DR). The training utilizes the Adam optimizer, a widely adopted choice in deep learning due to its adaptive learning rate capabilities and computational efficiency. The learning rate is set to 2×10^{-4}, a value determined empirically to balance convergence speed and stability. A mini-batch size of 4 is used during training, accommodating GPU memory limitations while allowing for frequent weight updates.

The models were trained to a total of 100 epochs, a timeframe adequate to achieve convergence and to avoid overfitting, according to validation performance. Categorical cross-entropy loss function is used, well-suited to multi-classification of mutually exclusive labels. Categorical cross-entropy calculates the difference between the observed label distribution and the predicted probability distribution of the five DR severities.

The training dataset is split into two subsets, whereby 90% of the dataset is reserved for training and 10% is saved for validation. The stratification split guarantees that every class is proportionally distributed in both subsets and maintains the class distribution to provide sound performance metrics when validating.

The subsequent step following independent backbone model training is meta-classifier construction and training. In validation set inference, every learned model gives rise to a five-dimensional probability vector (logits) of predicted DR class probabilities. The vectors, one by every model, are then concatenated to give a ten-dimensional feature vector to every image sample. This compound representation is input to the meta-classifier.

To avoid data leakage and ensure an unbiased estimation of performance, the meta-classifier is trained exclusively on the validation set. This approach ensures that the meta-classifier learns to combine the individual model outputs without being influenced by the training data seen by the base models. The meta-classifier itself is typically implemented as a shallow fully connected neural network, though other architectures may also be explored. It is trained to map the concatenated logits to the final DR classification label, optimizing a categorical cross-entropy loss.

This two-phase training strategy—first learning diverse representations from multiple EfficientNet variants and then combining them through a meta-classifier trained on unseen validation data—enhances the robustness and generalization capacity of the proposed ensemble framework.

D. Evaluation Metrics

Model performance is evaluated using several well-established metrics, each capturing different aspects of classification effectiveness and calibration.

IV. Evaluation Metrics

To assess the performance of the proposed model, several evaluation metrics were employed:

- Accuracy: Measures the proportion of correct predictions among the total number of samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

- F1 Score: Represents the harmonic mean of precision and recall, useful for handling class imbalances:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{Recall} = \frac{TP}{TP+FN}$.

- Cohen's Kappa: Evaluates the agreement between predicted and actual labels, correcting for chance agreement:

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

where p_o is the observed agreement and p_e is the expected agreement by chance.

- Area Under the Curve (AUC): Represents the area under the ROC curve, illustrating the classifier's ability to distinguish between classes:

$$\text{AUC} = \int_0^1 \text{TPR}(x)\, dx$$

where TPR is the true positive rate and FPR is the false positive rate.

- Brier Score: Measures the accuracy of probabilistic predictions as the mean squared difference between predicted probabilities and actual binary outcomes:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (f_i - o_i)^2$$

where f_i is the predicted probability and o_i is the actual outcome (0 or 1) for sample i.

- Expected Calibration Error (ECE): Evaluates the alignment of predicted probabilities with actual outcomes across multiple confidence bins:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

where M is the number of bins, B_m is the set of samples in bin m, $\text{acc}(B_m)$ is the accuracy, and $\text{conf}(B_m)$ is the average predicted confidence in that bin.

- Maximum Calibration Error (MCE): Indicates the worst-case deviation between confidence and accuracy across all bins:

$$\text{MCE} = \max_m \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

Lower MCE values denote better calibration.

This ensemble strategy offers improved classification accuracy and calibration, effectively leveraging the strengths of both EfficientNet variants. The proposed method demonstrated

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

superior performance over individual models, confirming that the stacking ensemble effectively integrates multi-scale features learned by EfficientNet-B3 and B4.

V. RESULTS

The synthetically balanced APTOS 2019 dataset was used in a number of experiments to assess the performance of the suggested ensemble framework. Many performance metrics, such as classification accuracy, class-wise precision, recall, F1-score, AUC, Brier loss, and calibration metrics like Expected Calibration Error (ECE) and Maximum Calibration Error (MCE), were used to evaluate the model. To assess inter-class agreement beyond chance, the Cohen's Kappa score was also calculated.

A. Classification Metrics

Table I presents the detailed classification performance across all five stages of Diabetic Retinopathy.Despite class imbalance, the ensemble model demonstrated balanced performance across all classes with an overall accuracy of 79% and a macro-averaged F1-score of 0.79.

TABLE I
CLASS-WISE EVALUATION METRICS ON VALIDATION DATASET

Class	Precision	Recall	F1-Score	Support
(No DR)	0.97	0.95	0.96	163
(Mild DR)	0.81	0.90	0.86	208
(Moderate DR)	0.70	0.60	0.64	172
(Severe DR)	0.70	0.75	0.72	191
(Proliferative DR)	0.77	0.73	0.75	176
Overall Accuracy	**0.79 (902 samples)**			
Macro Avg	0.79	0.79	0.79	902
Weighted Avg	0.79	0.79	0.79	902

B. AUC and Calibration

The Area Under the Curve (AUC) was computed per class, yielding the following scores:

AUC Class-wise = [0.9899, 0.9575, 0.8498, 0.9216, 0.9087]

The average AUC across all classes was 0.9255, indicating excellent discriminative capability.

Calibration quality was assessed using the Expected Calibration Error (ECE = 0.1319) and Maximum Calibration Error (MCE = 0.3558), reflecting the degree to which predicted probabilities align with actual outcomes. Additionally, the Brier loss was computed as 0.1039, reinforcing the model's strong calibration.

C. Agreement Score

A high degree of agreement between the model predictions and ground truth labels, beyond chance, was indicated by the validation set's Cohen's Kappa Score of 0.8939.

D. Comparison with Existing Methods

The performance of the suggested ensemble model is compared to current state-of-the-art methods for DR classification on the APTOS or comparable datasets in Table II. The suggested approach integrates uncertainty calibration and ensemble learning techniques that were not frequently covered in previous works, and it shows a noticeable improvement in both overall accuracy and macro F1-score.

TABLE II
COMPARING WITH THE LATEST DR DETECTION TECHNIQUES

Method	Accuracy (%)	F1-Score	Remarks
Kaggle DR Winner (2015)	75.0	0.71	Single CNN (Inception-v3)
Li et al. (2020)	77.6	0.74	Attention-based ResNet for retinal grading
Zhou et al. (2021)	78.4	0.76	CNN ensemble, no uncertainty calibration
Proposed Method(Ours)	**79.0**	**0.79**	EfficientNet-B3+B4 Ensemble

Key advancements of our model include:

- Use of a stacked ensemble of EfficientNet-B3 and B4 to integrate complementary feature hierarchies.
- Incorporation of probabilistic calibration (ECE, MCE, Brier loss), enabling more reliable uncertainty-aware pre- dictions.
- Extensive data balancing and augmentation to improve generalization to real-world clinical scenarios.

VI. DISCUSSION

A. Key Findings

The proposed ensemble-based deep learning model demonstrated robust performance in the multi-class classification of Diabetic Retinopathy (DR) stages. The use of EfficientNet-B3 and B4, fine-tuned on a synthetically balanced version of the APTOS 2019 dataset, contributed to improved classification metrics across all DR categories. Model generalization was further improved by data augmentation techniques like Gaussian blur and geometric transformations. The effectiveness of the suggested method for extensive screening applications is supported by evaluation metrics like accuracy (79%), macro-averaged F1-score (0.79), and Cohen's Kappa (0.89).

B. Clinical and Technical Implications

This study demonstrates that deep ensemble models, when properly trained and calibrated, can be deployed as scalable solutions for early-stage DR screening. The integration of diverse augmentation techniques helps simulate real-world variability in retinal fundus images, enhancing the model's robustness in practical settings. The proposed system holds significant potential for deployment in under-resourced regions where access to expert ophthalmologists is limited.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

VII. CONTRIBUTIONS AND IMPACT

The primary contributions of this work include:

- Development of an ensemble learning framework using EfficientNet-B3 and B4 for DR classification.
- Introduction of a synthetically balanced dataset derived from APTOS 2019 to mitigate class imbalance.
- Application of advanced augmentation and preprocessing techniques to simulate clinical diversity.
- Achieving better results than the state-of-the-art techniques, as shown by thorough evaluation metrics.

This research provides a concrete step toward the integration of AI-assisted diagnostics in ophthalmology, offering a foundation for clinical decision support systems.

VIII. LIMITATIONS

This study has some drawbacks:

- The dataset, while balanced synthetically, may not fully represent the distribution and diversity of real-world populations.
- External validation was limited to the APTOS 2019 dataset, restricting generalizability to other imaging en- vironments and demographics.
- The model, while calibrated, does not yet incorporate lesion-level localization, which is vital for interpretability in clinical workflows.

IX. FUTURE WORK

Building on the current findings, future research directions include:

- Incorporating more advanced ensemble strategies, such as weighted voting and Bayesian averaging, to further improve accuracy.
- Employing transfer learning from larger ophthalmic datasets and domain-specific pretraining for better feature extraction.
- Extending validation across multiple public and private datasets (e.g., Messidor, IDRiD) to assess model generalizability.
- Exploring explainable AI (XAI) techniques to visualize pathological regions contributing to DR stage predictions.

REFERENCES

[1] V. Gulshan et al., ”Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[2] M. D. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, ”Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning,” Investigative Ophthalmology Visual Science, vol. 57, no. 13, pp. 5200–5206, 2016.

[3] G. Quellec, K. Charrie, Y. Boudi, B. Cochener, and M. Lamard, ”Deep image mining for diabetic retinopathy screening,” Medical Image Analysis, vol. 39, pp. 178–193, 2017.

[4] R. Gargeya and T. Leng, ”Automated Identification of Diabetic Retinopathy Using Deep Learning,” Ophthalmology, vol. 124, no. 7, pp. 962–969, 2017.

[5] D. S. W. Ting et al., ”Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes,” JAMA, vol. 318, no. 22, pp. 2211–2223, 2017.

[6] T. Vo, A. La, T. Nguyen, and D. Pham, ”A Novel Deep Learning Approach for Diabetic Retinopathy Detection on Imbalanced Data,” in Proceedings of the 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1–6.

[7] K. K. Maninis, J. Pont-Tuset, P. Arbela´ez, and L. Van Gool, ”Deep Retinal Image Understanding,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016, pp. 140–148.

[8] G. Quellec, M. Lamard, P. Josselin, G. Cazuguel, B. Cochener, and C. Roux, ”Optimal wavelet transform for the detection of microaneurysms in retina photographs,” IEEE Trans. Med. Imaging, vol. 27, no. 9, pp. 1230–1241, 2008.