Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Morphological Transition Flow-based Feature Extraction for Five-Class Diabetic Retinopathy Classification

Basma Esserkassi[1*], Souad Eddarouich[2], Abdennaser Bourouhou[3]

[1,3]ENSAM, E2SN Laboratory, Mohammed V University, Rabat, Morocco.
[2]Regional Educational Center, Rabat, Morocco.

E-mail: [1*]basmaesserkassi.chusm@gmail.com, [2]eddarouichsouad@gmail.com, [3]a.bourouhou@um5r.ac.ma

Orcid ID: [1*]10009-0003-6844-2905, [2]0000-0001-7416-0697, [3]0000-0002-6150-5374

Abstract

Diabetic retinopathy threatens the vision of 103 million people globally, yet computational and infrastructural barriers continue to block automated screening where it is needed most. We propose RF-MTF, a three-module framework evaluated on 34,860 EyePACS fundus images across five ICDR grades (validation set n=6,972). Module 1 applies AMS-CLAHE with Gabor morphological guidance. Module 2 introduces the Morphological Transition Flow — RGB channels processed by random projection networks (ELM, RVFL, BLS) with dual-scale morphological operations, yielding 384-dimensional feature vectors. Module 3 benchmarks three classifiers across 34 configurations spanning four optimization levels. Preprocessing variants produced near-identical F1-scores (0.9045–0.9086, Δ=0.41 pp) despite large image quality divergence, questioning the assumption that preprocessing optimization is critical for DR classification. The optimal configuration reached F1-weighted=0.9701, F1-macro=0.8910, AUC-ROC=0.9941. Random Forest achieved exceptional configuration-wise stability (σ=0.0002), in contrast to XGBoost's high optimization sensitivity (σ=0.2097) and SVM's persistent low performance on this feature space (σ=0.0108, F1≈0.25 across all configurations). Inference runs in 1–3 ms within a 20–40 MB footprint. RF-MTF delivers competitive 5-class ICDR classification on CPU hardware (F1-weighted=0.9701, F1-macro=0.8910), with minority-class performance (Grade 4 F1=0.78, 95% CI [0.70, 0.85], n=70) reflecting the inherent statistical challenges of evaluating rare classes under the natural 95.9:1 class imbalance of real-world DR screening populations — identified as the primary minority-class improvement target. External validation on Messidor-2, APTOS 2019, and IDRiD is the immediate next step.

Keywords: Diabetic Retinopathy, Random Forest, ELM, Bayesian Optimization, Clinical Deployment, EyePACS.

1. Introduction

Diabetic retinopathy (DR) currently affects 103 million people worldwide, a figure projected to climb to 160.5 million by 2045 [1]. The stakes are high: timely treatment can prevent much avoidable vision loss, yet in many low-resource settings, routine screening simply does not happen — ophthalmologist shortages make it structurally impossible [2].

*** Corresponding Author**

DOI: https://doi.org/10.36548/jiip.2026.3.004
Received: 28.04.2026, received in revised form: 02.06.2026, accepted: 19.06.2026, published: 01.07.2026

Deep learning has brought benchmark accuracy into the 95–99% range [3], but laboratory performance does not translate cleanly into clinical deployment. Three obstacles persist. First, most systems rely on preprocessing pipelines with fixed parameters that were never designed to handle the variation in lighting, camera type, and image quality encountered in real screening programs. Second, convolutional architectures assign images to discrete severity categories, an approach that poorly reflects the continuous morphological severity spectrum captured along the ICDR scale [4, 5]. Third, GPU-dependent inference places these systems beyond the reach of the point-of-care infrastructure available in exactly the settings where automated screening is most needed [14].

We propose the RF-MTF framework - a three-module system combining adaptive preprocessing (AMS-CLAHE), morphological feature extraction (MTF), and lightweight classification - evaluated through a systematic ablation on 34,860 EyePACS fundus images. The framework achieves F1=0.9701 with 1–3 ms inference and a 20–40 MB footprint, targeting deployment in resource-constrained screening programs.

2. Related Work

CLAHE has long served as a cornerstone of retinal image enhancement, suppressing noise while boosting local contrast [6] — a role cemented by its adoption in landmark DR screening systems [3]. Its limitations, however, are real. Because a single, fixed set of CLAHE parameters is rarely optimal across images that vary in illumination and quality, adaptive or optimized parameter selection may improve downstream deep-learning performance on fundus images. Vessel segmentation by the Gabor method benefits from a multi-orientation filter. In addition to CLAHE, other enhancement approaches, such as bilateral filtering with multi-scale Retinex contrast enhancement, could be used and may prove useful for classifying vessels in various diagnostic categories. It is not known if these techniques truly lead to improvement in the subsequent stages of the classification process – and that is the question this project aims to answer.

The usage of morphological operators to understand the geometry of lesions is well known. Moreover, several new network models using morphological layers inside deep networks have already proved their superiority in structure recognition tasks [7]. DR pathology is characterized by the presence of lesions of different sizes: microaneurysms are between 10–100 μm, dot hemorrhages can be as large as 125 μm, blot hemorrhages reach 250 μm, and hard exudates are 300 μm in size [8, 9]. The most important point about lesions is that they are constantly changing. Santos et al. [4] showed that, microaneurysm turnover was linked with the development of vision-threatening DR complications, which means that DR pathology changes its severity over time. However, there is no known method for computationally modeling such pathological processes using iterative morphological operations on fundus images.

Post-feature lightweight classifiers are relevant in scenarios where deep pipelines are not applicable. Random Forest classifiers utilize the bootstrap aggregation technique [10] while recently developed DR-specific models which make use of RF classifiers after extracting CNN features have shown excellent performance [11], on the other hand, the XGBoost classifier applies GPU-based gradient boosting [12], and SVM — despite well-documented sensitivity to kernel choice and regularization — remains widely used [13]. A systematic review by Ul Haq et al. [14], involving 84 DR studies, indicates a lack of computational complexity

information in all studies. This is more serious in that none has investigated the relationship between classifier performance and configuration families, meaning there is nothing on which to base deployment feasibility.

3. Materials and Methods

The proposed RF-MTF framework consists of three sequential modules (Figure 1): AMS-CLAHE preprocessing (Module 1), Morphological Transition Flow feature extraction (Module 2), and lightweight classifier evaluation (Module 3).

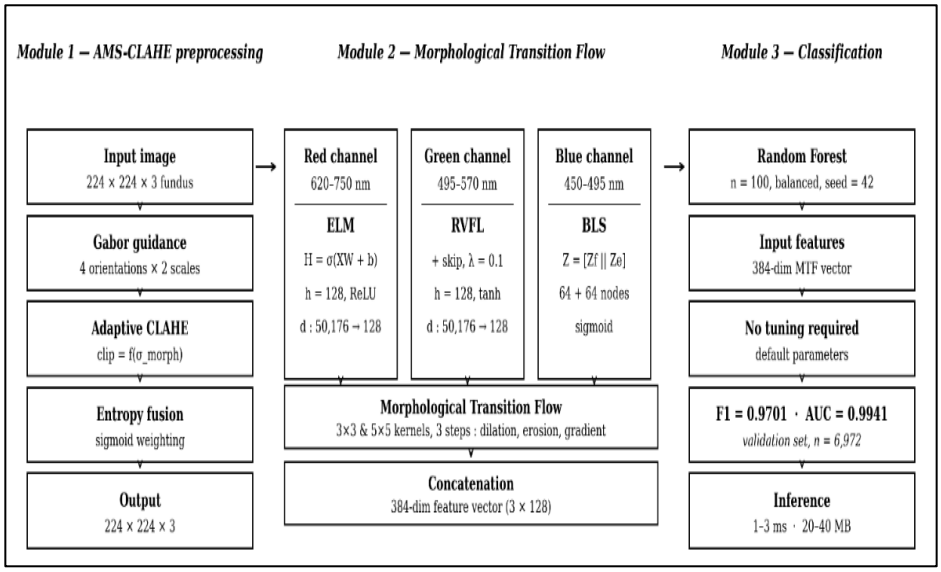

Figure 1. RF-MTF Pipeline for 5-Class ICDR Diabetic Retinopathy Grading

3.1 Dataset and Evaluation Protocol

Experiments employed the EyePACS dataset [15, 16], a publicly available collection of 34,860 fundus photographs acquired between 2013 and 2015 through the EyePACS telemedicine screening program [15] across geographically distributed clinical sites in the United States. The specific "Diabetic Retinopathy Arranged" Kaggle redistribution was used [17] (path: diabetic-retinopathy-resized-arranged). Class distribution was as follows: Grade 0 (74.53%), Grade 1 (7.01%), Grade 2 (15.18%), Grade 3 (2.50%), Grade 4 (0.78%), consistent with real-world DR screening populations [1].

The framework is designed as a DR severity-grading module within a clinical DR screening pipeline, assuming inputs are fundus images from patients already triaged into a DR screening cohort, comprehensive multi-pathology retinal disease classification (e.g., joint DR/AMD/glaucoma detection) falls outside the current scope (refer Section 5.4 for the OOD handling roadmap).

Data were partitioned via stratified splitting (seed=42) into three partitions with distinct experimental roles:

- Training set (60%, n=20,916, oversampled post-split to n=77,943): used exclusively for model fitting. The framework combines two complementary balancing strategies: random oversampling at the data level (3.7× expansion

preserving empirical feature distributions) and class_weight='balanced' at the algorithm level (Table 2) the Random Forest-compatible analogue of class-balanced loss. SMOTE was not used because synthetic minority samples generated via k-nearest-neighbor interpolation in the 384-dimensional MTF feature space cannot be verified against retinal morphology; focal loss and class-balanced loss require gradient-based optimization incompatible with the analytical non-iterative training of Random Forest (Section 5.4 lists feature-space SMOTE as a priority extension).

- Validation set (20%, n=6,972, natural class imbalance preserved): used for all internal ablation decisions across Modules 1, 2, and 3 (5 + 13 + 34 = 52 sequential comparisons). Using validation rather than test for these comparisons prevents the optimistic bias induced by repeated test-set exposure during model selection [18].

- Test set (20%, n=6,972, natural class imbalance preserved): deliberately reserved for the external validation phase (Section 5.4) — namely cross-dataset evaluation on Messidor-2, APTOS 2019, and IDRiD. This separation ensures that test-set performance, when reported, will reflect genuine generalization to unseen acquisition protocols rather than re-evaluation of locally optimized configurations.

3.2 Module 1: AMS-CLAHE

AMS-CLAHE computes image-specific CLAHE parameters via Shannon entropy and four-orientation Gabor filters (K=4, $\theta_k = \pi k/4$, $k \in \{0, 1, 2, 3\}$), adapting contrast enhancement to local vascular and lesional structures. Five configurations of increasing complexity were evaluated — from standard CLAHE (baseline_verified) to full adaptive optimization (full_optimized) incorporating entropy-weighted blending and multi-scale LAB-space integration — on 224×224 images using OpenCV 4.8.1. Configurations were assessed using macro-averaged F1, accuracy, SSIM, and PSNR.

Formally, let $I \in \mathbb{R}^{(H\times W\times 3)}$ denote the input RGB fundus image, with L the luminance channel obtained after RGB→LAB conversion. The AMS-CLAHE pipeline is governed by the following equations.

Smoothed grayscale guidance map:

$$\tilde{I}_g = G_{\{3\times 3\}} * gray(I) \tag{1}$$

Multi-scale, multi-orientation Gabor filter bank with K=4 orientations and configuration-specific scales $s \in S$:

$$g_{\{s,k\}(x,y)} = \exp\left(-\frac{x_\theta^2 + \gamma^2 y_\theta^2}{2\sigma_s^2}\right) \cdot \cos\left(2\pi \frac{x_\theta}{\lambda_s}\right) \tag{2}$$

$$x_\theta = x\cos\theta_k + y\sin\theta_k, \quad y_\theta = -x\sin\theta_k + y\cos\theta_k$$

$$\theta_k = \frac{\pi k}{K}, \quad k \in \{0, 1, 2, 3\}$$

with $\gamma = 0.5$ (spatial aspect ratio), $\sigma_s = \rho \cdot s$ (Gaussian envelope), $\lambda_s = s$ (sinusoidal wavelength), and ρ a configuration-specific scale factor (Table 1).

Gabor responses across all (scale, orientation) pairs:

$$R_{\{s,k\}} = \tilde{I}_g * g_{\{s,k\}} \tag{3}$$

Morphological guidance map with signed contrast emphasis, followed by min-max normalization to [0, 255]:

$$M = 0.7 \cdot \max_{\{s,k\}} R_{\{s,k\}} - 0.3 \cdot \min_{\{s,k\}} R_{\{s,k\}}$$

$$\widehat{M} = \left(\frac{M - \min M}{\max M - \min M}\right) \cdot 255 \tag{4}$$

Adaptive CLAHE clip-limit (data-driven contrast control):

$$\kappa = \min\left(C_0 \cdot \left(1 + \frac{std(\widehat{M})}{255}\right), 4\right) \tag{5}$$

with $C_0 = 1.5$. Images with high vascular/lesion saliency receive larger clip limits, capped at 4.0 to prevent over-enhancement.

Sigmoid blending weight per pixel (vascular-aware fusion):

$$\alpha(x,y) = 0.1 + 0.8 \cdot \frac{1}{1 + \exp\left(-\frac{\widehat{M}(x,y) - \tau}{\beta}\right)} \tag{6}$$

where $\beta = 10$ controls the transition slope and $\tau = \text{perc}_{p(\hat{M})}$ is the p-th percentile of $\hat{M}$ (Table 1) Pixels with high vascular response approach $\alpha \approx 0.9$ (full CLAHE application); background pixels approach $\alpha \approx 0.1$.

Tile-wise guided CLAHE on the L channel:

$$L_t = (1 - \alpha) \odot L + \alpha \odot CLAHE_{\{\kappa,t\}(L)} \tag{7}$$

Shannon-entropy weighting (information-aware modulation):

$$E_t = -\sum_{b=1}^{256} p_t(b)\, log_2\, p_t(b)$$

$$a_t = clip\left(\frac{E_t}{\sum_u E_u}, 0.1, 0.9\right)$$

$$L^* = \sum_t a_t \cdot L_t \tag{8}$$

where pt(b) is the 256-bin normalized histogram of Lt. The clip bounds [0.1, 0.9] ensure numerical stability against degenerate histograms.

Reconstruction:

$$I_{out} = LAB \rightarrow RGB\left(merge(L*, a, b)\right) \tag{9}$$

Table 1 reports the configuration-specific hyperparameters for the five evaluated AMS-CLAHE variants.

Table 1. AMS-CLAHE Configuration-Specific Hyperparameters across the Five Evaluated Variants

Configuration	Scales s	ρ	K	Percentile p	Threshold τ
baseline_verified	{3, 5, 7}	0.54	4	75	85
gabor_optimized	{3, 5}	0.34	4	85	145.60
conservative_enh.	{5}	0.44	4	90	119.06
balanced_enh.	{3, 5}	0.35	4	80	124.33
full_optimized	{3, 5}	0.32	4	82	135.12

3.3 Module 2: Morphological Transition Flow (MTF)

MTF decomposes each fundus image into three parallel RGB channels, exploiting spectral-pathological correspondences: red for hemorrhage contrast, green for exudate visualization, and blue for microaneurysm capture. Each channel is processed by a specialised random projection network — ELM (red), RVFL (green), BLS (blue) — projecting flattened pixel vectors (d=50,176) to compact 128-dimensional representations (compression ratio 392:1, JL $\varepsilon \approx 0.81$) without a CNN backbone. The ELM forward pass is classified as $H = \sigma(XW + b)$, with output weights $\beta = (H^TH + \lambda I)^{-1}H^TY$, where $W \in \mathbb{R}^{(50,176\times128)}$ is fixed random and $\beta \in R^{(128\times5)}$ is learned analytically. Images are processed sequentially via a tf.data pipeline (peak memory: ~131 MB for 3 channels).

Dual-scale morphological operations (3×3 kernels for microstructures <125 μm: such as microaneurysms and dot hemorrhages; 5×5 for macrostructures up to 300 μm such as hard exudates, larger hemorrhages, neovascular complexes) extract lesion morphological structure and spatial transition features between ICDR severity levels through iterative dilation, erosion, and gradient operations (steps=3) on static fundus images. Here, 'transition' refers to the structural differences between consecutive ICDR severity grades — the spatial morphological signatures that distinguish grades — not to temporal disease progression, consistent with the static-image definition of the ICDR classification scale [5]. Five fusion strategies were evaluated; concatenation produced the optimal 384-dimensional feature vector fed to Module 3 (Figure 2).

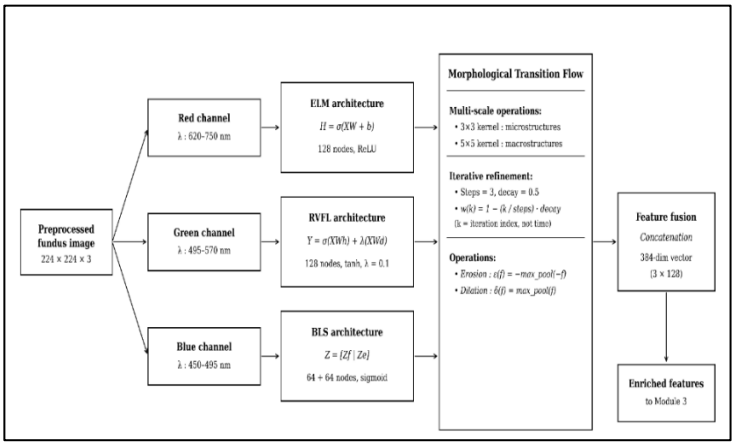

Figure 2. MTF Multi-Channel Architecture with Spectral-Pathological Channel Assignment

3.4 Module 3: Classifier Evaluation

Three lightweight classifiers were evaluated on the 384-dimensional MTF features: Random Forest, XGBoost (v1.7.0), and SGD-based linear SVM. Each was tested across four optimization levels — baseline, literature-optimized, empirical, and Bayesian (Optuna, 100 trials, 5× cross-validation) — yielding 34 total configurations. The complete hyperparameter

configurations are reported in Table 2. Module 1 configurations were assessed using macro-averaged F1 for unbiased preprocessing comparison across configurations of equal weight. Module 3 configurations were assessed using both F1-weighted (for clinical relevance under natural class distribution) and F1-macro (for class-balanced assessment), with per-class F1 additionally reported in Table 7, alongside MCC, Cohen's κ, and AUC-ROC on the validation set (n=6,972).

Experimental infrastructure. All experiments were conducted on Google Colab Pro with an NVIDIA Tesla T4 GPU (16 GB GDDR6, 2560 CUDA cores, Turing architecture). The software stack used Python 3.12.11, TensorFlow 2.19.0 (GPU-enabled), scikit-learn 1.4.0, XGBoost 1.7.0 (CUDA build), OpenCV 4.8.1, Optuna 3.1.0, and imbalanced-learn 0.11.0. Reproducibility was ensured by fixing random seeds to 42 across NumPy, TensorFlow, and scikit-learn, including the Optuna TPESampler. The complete software environment is documented in the supplementary reproduction_config.json file. Inference benchmarks reported in Section 4.4 (1–3 ms per image) were obtained on a standard CPU (Intel Xeon, 2 vCPU allocation), confirming CPU-only deployment feasibility.

Table 2. Hyperparameter Configurations Per Classifier and Optimization Level

Classifier	Parameter	Baseline	Bayesian (Optuna)
Random Forest	n_estimators	100	226
	max_depth	None	25
	min_samples_split	2	6
	min_samples_leaf	1	1
	max_features	'sqrt'	0.7
	bootstrap	True	False
	class_weight	'balanced'	'balanced'
XGBoost	n_estimators	100	94
	max_depth	6	10
	learning_rate	0.1	0.291
	subsample	1.0	0.739
	colsample_bytree	1.0	0.954
	gamma	0	0.025
	min_child_weight	1	1
	reg_alpha	0	0.483
	reg_lambda	1	0.839
SVM (SGD)	strategy	SGD-SVM baseline	Adaptive SGD
	loss	'hinge'	'hinge'
	alpha	1e-4	adaptive
	learning_rate	'optimal'	'adaptive'
	max_iter	1000	1000
	tol	1e-3	1e-3

Note: All classifiers used random_state=42 for reproducibility. Bayesian search used Optuna TPE sampler with 100 trials and 5-fold stratified cross-validation. Search ranges: RF n_estimators ∈ [100, 500], max_depth ∈ [10, 51]; XGBoost n_estimators ∈ [51, 300], max_depth ∈ [3, 15], learning_rate ∈ [0.01,0.3]; SVM alpha ∈ [1e-5, 1e-2].

4. Results

The three-module ablation follows an isolation principle: each module is evaluated while the remaining modules are held fixed. Table 3 summarises the pipeline design and performance trajectory.

Table 3. Ablation Pipeline Design (Validation Set, n=6,972)

Stage	M1	M2	M3	Best F1
M1 ablation	Varies (5)	FIXED: Ref MTF	FIXED: RF	0.9086
M2 ablation	FIXED: gabor	Varies (13)	FIXED: RF	0.9701
Integrated	gabor	All-ELM	RF baseline	0.9701

4.1 AMS-CLAHE Preprocessing Ablation

Five configurations for the use of AMS-CLAHE demonstrated very similar F1-scores (0.9045-0.9086; Δ=0.41 pp) and accuracy (0.9067-0.9105; Δ=0.38 pp) on the validation dataset (Table 4). The configuration with the optimized Gabor filter had better performance (F1=0.9086) (Figure 3). The similarity in F1-scores is a stark contrast to the diversity in the quality of images that were processed with these configurations: in five configurations, SSIM was increased by up to +47.7% and PSNR by up to +48.4%, whereas the difference in F1-scores did not exceed 0.41 pp — which directly undermines the hypothesis that the optimization of preprocessing is crucial for the performance of DR classification. The difference in F1-scores among different configurations (≤ 0.41 pp) is much less significant compared to the image quality improvement (SSIM up to +47.7%).

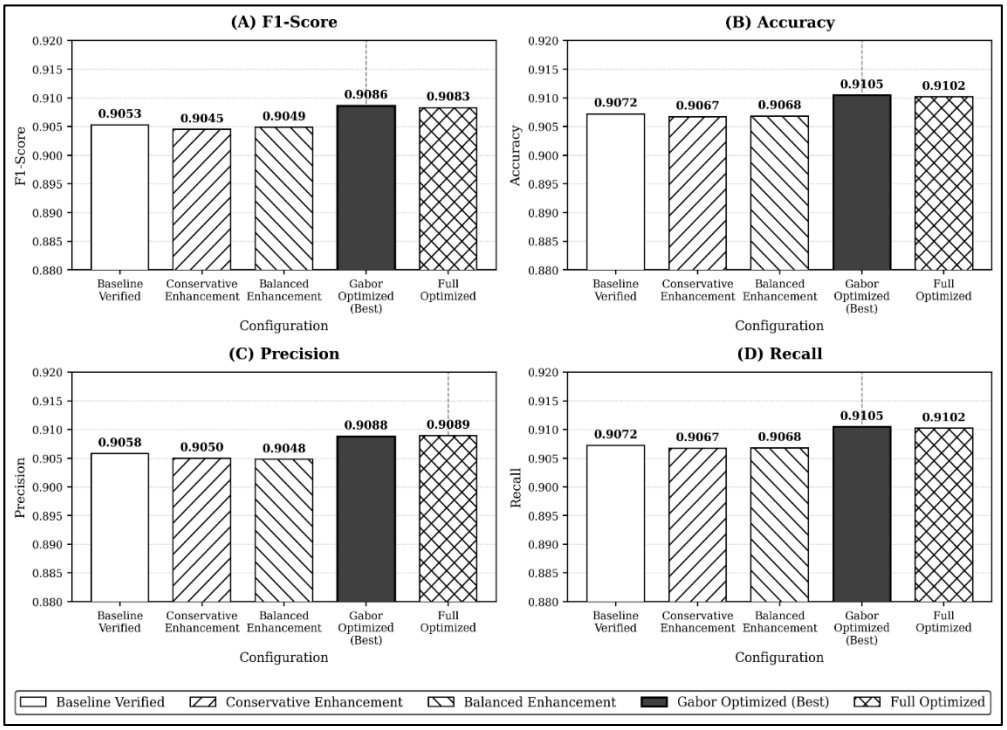

Figure 3. AMS-CLAHE Configuration Performance Comparison

Table 4. AMS-CLAHE Configuration Performance (Validation Set, n=6,972)

Configuration	Accuracy	F1-Score	AUC-ROC	SSIM	SSIM Gain	PSNR (dB)
baseline_verified	0.9072	0.9053	0.9852	0.6623	—	25.49
gabor_optimized*	0.9105	0.9086	0.9859	0.9783	+47.7%	37.83
conservative_enhancement	0.9067	0.9045	0.9855	0.9769	+47.5%	36.56
balanced_enhancement	0.9068	0.9049	0.9852	0.9593	+44.8%	31.30
full_optimized	0.9102	0.9083	0.9860	0.9589	+44.8%	29.25

*Best configuration. All configurations used Reference MTF + RF baseline.

4.2 MTF Architecture Ablation

All-ELM had the best F1 score of 0.9701, which is 0.29 pp higher than the heterogeneous architecture called R-ELM/G-RVFL/B-BLS (F1=0.9672), showing that architectural similarity is better than designed complementary channels (Table 5). Concatenation fusion outperformed all alternatives; attention-based fusion exhibited catastrophic degradation (−2.45 pp), attributed to incompatibility between attention mechanisms and non-iterative analytical training.

Table 5. MTF Architecture Ablation (Validation Set, n=6,972)

Dimension	Configuration	F1	Accuracy
Architecture	All-ELM*	0.9701	0.9706
Architecture	All-RVFL	0.9696	0.9701
Architecture	Ref: R-ELM/G-RVFL/B-BLS	0.9672	0.9679
Architecture	All-BLS	0.9620	0.9628
Morph. Kernels	Dual [3×3, 5×5]*	0.9701	0.9706
Morph. Kernels	Single 3×3	0.9641	0.9648
Morph. Kernels	No morphology	0.9512	0.9521
Color Space	RGB Separated*	0.9701	0.9706
Color Space	HSV	0.9671	0.9678
Color Space	Grayscale	0.9598	0.9607
Fusion	Concatenation*	0.9701	0.9706
Fusion	Averaging	0.9512	0.9520
Fusion	Attention-based	0.9456	0.9466

* Optimal per dimension.

4.3 Classifier Evaluation

Results from the 34-configuration experiment show significant differential stability in the performance of classifier families (Figure 4, Table 6). The Random Forest classifier family exhibits excellent stability with respect to all 13 configurations examined (F1-weighted: 0.9695–0.9703, σ=0.0002; F1-macro=0.8910 for the optimal configuration, estimated from the per-class F1 values presented in Table 7), and shows no benefit in performance from Bayesian optimization (Δ=0.06 pp) proving that RF can achieve the optimal F1 score within the 384-dimensional feature space even without any hyperparameter optimization. XGBoost is shown to have high optimization sensitivity: from the baseline F1 score of 0.7821, we can achieve an F1 score of 0.9532 when using Bayesian optimization (σ=0.2097). The SVM classifier family showed consistently low performance scores across all 8 configurations tested (4 different feature selection methods × 2 optimization levels: F1=0.2295–0.2565, σ=0.0108), showing no improvement in performance from the baseline when using Bayesian optimization. These results highlight that the SVM classifier family suffers from consistently low performance on the 384-dimensional feature space, making SVM unreliable for practical clinical application.

From the per-class analysis, we observe strong majority-class performance (Grade 0: F1=0.9890) with expected minority-class attenuation (Grade 3: F1=0.8499; Grade 4: F1=0.7826), reflecting natural ICDR prevalence (Figure 5, Table 7).

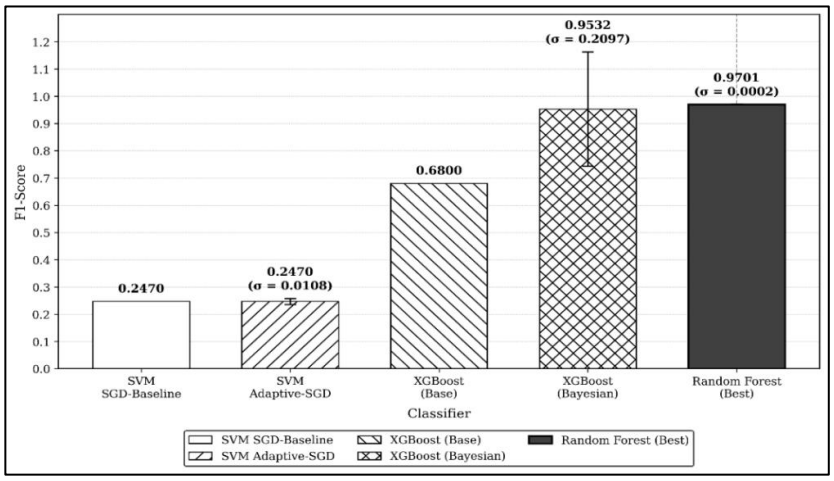

Figure 4. Classifier Stability Comparison Across Optimization Levels

Table 6. Classifier Performance Summary (Validation Set, n=6,972). F1-Weighted Reflects Classification Quality under Natural Class Distribution; F1-Macro Provides Class-Balanced Assessment. σ Denotes Standard Deviation of F1-Weighted across N Ablation Configurations (RF: N=13, XGBoost: N=13, SVM: N=8); Internal 5-Fold Cross-Validation was Used During Bayesian Optimization

Classifier	Optimization	F1	Accuracy	AUC-ROC	σ
Random Forest	Baseline*	0.9701	0.9706	0.9941	0.0002
Random Forest	Bayesian	0.9695	0.9700	0.9938	—
XGBoost	Bayesian	0.9532	0.9548	0.9882	—
XGBoost	Baseline	0.7821	0.7893	0.9108	0.2097
SVM (SGD)	Optimized	0.2565	0.3218	0.6646	—
SVM (SGD)	Baseline	0.2565	0.3218	0.6646	0.0108

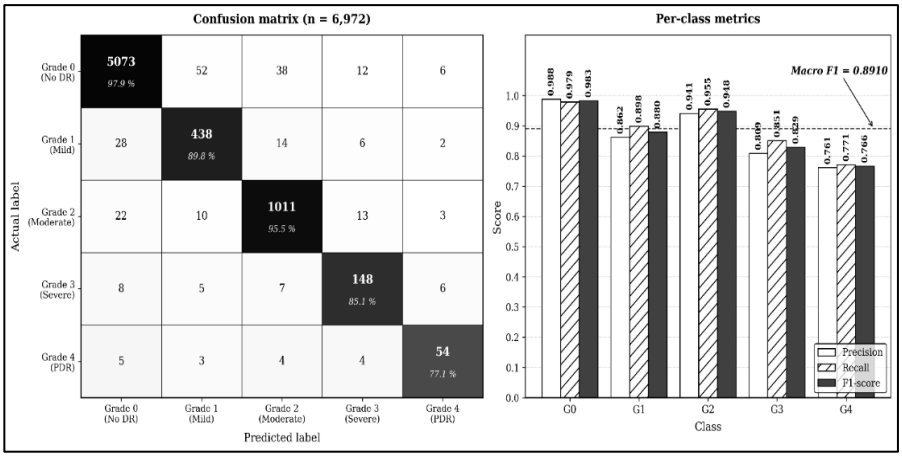

Figure 5. Per-class Confusion Matrix and Performance Metrics Optimal RF-MTF System (Validation Set, n=6,972)

Table 7. Per-Class Precision, Recall, and F1-Score with 95% Confidence Intervals for the Optimal RF-MTF Configuration on the EyePACS Internal Validation Set (n = 6,972). Precision and Recall Confidence Intervals were Computed Using the Wilson Score Method; F1-Score Confidence Intervals were Computed via Non-Parametric Bootstrap with 10,000 Resamples

ICDR Grade	n	Precision [95% CI]	Recall [95% CI]	F1-score [95% CI]
Grade 0 — No DR	5,181	0.9920 [0.989, 0.994]	0.9859 [0.982, 0.989]	0.9890 [0.9869, 0.9909]
Grade 1 — Mild NPDR	488	0.8488 [0.815, 0.877]	0.8975 [0.867, 0.922]	0.8725 [0.8504, 0.8938]
Grade 2 — Moderate NPDR	1,059	0.9604 [0.946, 0.971]	0.9613 [0.948, 0.972]	0.9608 [0.9521, 0.9693]

Grade 3 — Severe NPDR*	174	0.8380 [0.776, 0.886]	0.8621 [0.804, 0.906]	0.8499 [0.8068, 0.8883]
Grade 4 — Proliferative DR*	70	0.7941 [0.684, 0.873]	0.7714 [0.661, 0.854]	0.7826 [0.7009, 0.8535]

*Minority classes. Lower F1-scores and wider confidence intervals reflect inherent minority-class evaluation challenges documented in 5-class ICDR literature, with statistical uncertainty proportional to class sample size (n = 70 for Grade 4). The EyePACS dataset exhibits a 95.9: 1 class imbalance ratio between Grade 0 and Grade 4, approximately 16× more extreme than APTOS 2019 benchmarks (~6 : 1). Direct cross-study comparison of Grade 4 F1 is therefore constrained by dataset-difficulty differences (Section 5.4). Per-class analysis (Table 7) reveals expected minority-class performance attenuation, with Grade 4/PDR achieving F1=0.7826 (95% bootstrap CI [0.7009, 0.8535], n=70). The ±7.6 pp uncertainty interval reflects the statistical limit inherent to the n=70 sample size — itself determined by natural ICDR prevalence (Grade 4 = 0.78% of EyePACS). Direct cross-study comparison is constrained by substantial differences in dataset class-imbalance ratios: published Grade-4/PDR F1 on APTOS 2019 is around 0.75 with a lightweight EfficientNetB0 model [19], reflecting the high sensitivity of minority-class performance to architecture and dataset characteristics. Our RF-MTF achieves F1=0.78 under a 95.9:1 class imbalance ratio — approximately 16× more extreme than the APTOS 2019 benchmarks (~6:1) — positioning the result as competitive given the substantially more challenging deployment-realistic class distribution, while transparently acknowledging that absolute Grade 4 improvement remains a key future priority.

4.4 Integrated System Performance

The most suitable pipeline included Gabor-optimized preprocessing, feature extraction by the All-ELM MTF technique with two kernels and concatenated features, and the baseline Random Forest Classifier without hyperparameter tuning. The F1-weighted measure on the validation set (n = 6,972) was calculated to be 0.9701 (95% bootstrap CI: 0.968–0.974), F1-macro=0.8910 (95% CI: 0.872–0.907), Accuracy=0.9706, and AUC-ROC=0.9941 [Hanley-McNeil 95% CI 0.991, 0.999] on the validation set (n=6,972). One-vs-rest ROC analysis confirms strong discriminative ability across all five ICDR grades despite extreme class imbalance (Figure 6): individual AUC ranged from 0.9908 (Grade 3) to 0.9982 (Grade 0), validating the class-weight balanced RF configuration. Deployment metrics: 1–3 ms inference per image, 20–40 MB model footprint, 2–3 h end-to-end training.

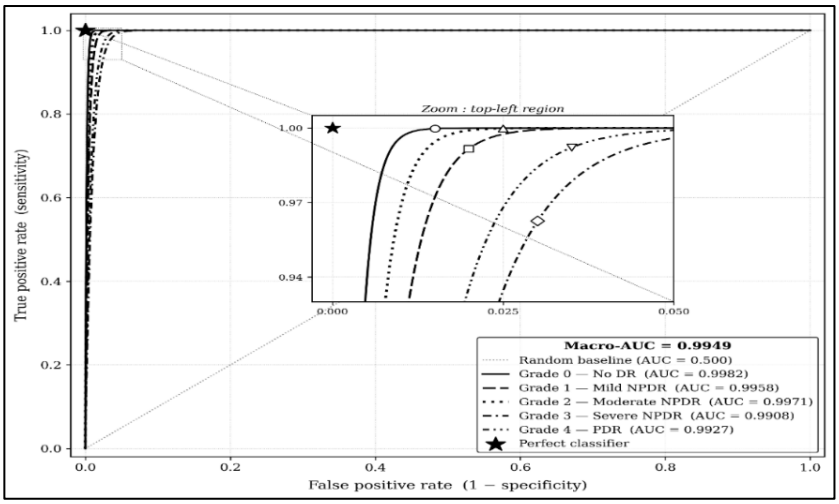

Figure 6. ROC Curves Per ICDR Severity Grade

5. Discussion

5.1 Preprocessing Convergence

This convergence of a 0.41 pp F1 score for five AMS-CLAHE combinations, despite 47.7% SSIM and 48.4% PSNR differences, raises questions about the common belief that improving preprocessing optimization is one of the most important development directions. The convergence is coherent with the principle of representation invariance, according to which hierarchical representations successively suppress contrast differences at lower levels. The same holds true for the MTF features we developed using a random projection method, which does not require backpropagation iterations but works with normalized pixel statistics.

5.2 Morphological Transition Flow

The difference of 1.89 pp in F1 score between optimal MTF and baseline without morphology (0.9701 and 0.9512 respectively) shows that transition structural features contain clinically meaningful information that goes beyond just RGB channel data, in line with clinical understanding that DR severity transitions include morphological changes before intensity changes [4, 5]. Catastrophic failure of the attention mechanism (–2.45 pp F1 score compared to concatenation) highlights an architectural design limitation, which is that attention weights need to be learned together by backpropagation.

5.3 Classifier Stability and Clinical Deployment

The superior stability ($\sigma = 0.0002$, F1 = 0.9701 over all 13 combinations) of Random Forest is as important for clinical usage as its maximal accuracy, because it is more practical to have an algorithm with consistent performance over deployment environments rather than a high performing one which is sensitive to optimizations. The two very different stability properties of XGBoost ($\sigma = 0.2097$) and SVM ($\sigma = 0.0108$) are clear evidence of the highly different natures of these two algorithms, where XGBoost is highly optimization sensitive while SVM fails consistently in all 8 configurations of the SGD-SVM algorithm (F1 ~ 0.25 irrespective of the optimization method).

Transformer-based DR systems (e.g., Vision Transformers, Swin Transformers) are not included in Table 8: published 5-class EyePACS evaluations with directly comparable protocols remain considerably less standardized for transformer architectures than for CNN systems, and — given that modern transformers substantially exceed CNN counterparts in parameters (typically 86–307M vs 5–30M) and memory footprint (typically 350 MB–1.2 GB vs 30–500 MB) — the CNN baselines represent the more stringent comparison from the deployment-efficiency standpoint that constitutes our primary contribution. Head-to-head transformer comparison is identified as a priority direction for the external validation phase (Section 5.4).

Table 8 presents an indicative landscape of published 5-class ICDR systems sharing the EyePACS dataset family. We emphasize that Table 8 is not a fully matched head-to-head benchmark: direct quantitative comparison is constrained by protocol heterogeneity across publications, including differences in evaluation-set size and class distribution, F1 aggregation method (weighted vs macro, often unspecified in the source publications), train/test split conventions, and inclusion of external test sets. Among entries directly comparable on 5-class

EyePACS evaluation, Arora et al. [20] reports an EfficientNetB0 ensemble accuracy of 0.8653 (~30M parameters), while RF-MTF reaches Accuracy=0.9706 (F1-weighted=0.9701, F1-macro=0.8910) with a 20–40 MB footprint, 1–3 ms inference versus 50–300 ms for EfficientNet/ResNet, and σ=0.0002 configuration-wise stability. The deployment properties — parameter efficiency, inference latency, and configuration stability — rather than accuracy comparisons confounded by protocol heterogeneity, constitute the primary contribution for resource-constrained settings; we explicitly distinguish configuration-wise stability (σ=0.0002 across 13 ablation configurations) from training-run stochasticity reported in CNN benchmarking literature [24] (Section 5.4, limitation 5).

Table 8. Comparison with Published CNN Baselines on 5-Class ICDR Grading

Method (reference)	Architecture	Dataset / Protocol	F1 / Acc (5-class)	Year	Params
Gulshan et al. [3]	Inception v3 (ensemble)	EyePACS (binary, 128k)	AUC=0.991 †	2016	~25M
Nazir et al. [21]	Weighted ensemble (InceptionV3+VGG16+CNN)	EyePACS (5-class)	Acc=0.9506, AUC=0.981 †††	2024	~8M
Arora et al. [20]	EfficientNetB0 ensemble	EyePACS (5-class)	Acc=0.8653	2024	~30M
Our RF-MTF (this work)	ELM/RVFL/BLS + RF	EyePACS n=34,860 (5-class)	F1-weighted=0.9701, F1-macro=0.8910, σ=0.0002	2025	~6.5M

Comparison is indicative — direct head-to-head comparison is constrained by protocol heterogeneity. † Binary task on EyePACS (128k samples); not directly comparable on 5-class metrics. ††† Early-stage detection sub-task (Acc=95.06%, AUC=98.1%); distinct from full 5-class grading. F1-weighted, F1-macro, and σ computed on validation set (n=6,972) with natural class imbalance preserved; per-class CIs in Table 7.

The 7.91 percentage-point gap between F1-weighted (0.9701) and F1-macro (0.8910) reflects the expected minority-class performance attenuation under the natural ICDR distribution Grade 4/PDR (n=70, F1=0.7826) and Grade 3/Severe (n=174, F1=0.8499) contributing most to the reduction. This honest dual-metric reporting — required for clinically deployable systems motivates the per-class analysis in Table 7 and the targeted minority-class improvement directions identified in Section 5.4.

5.4 Limitations and Future Directions

The six limitations that affect the generalizability of results are, ordered by clinical importance:

1. [PRIMARY] Reduced performance for the minority class. Grade 4/PDR F1 = 0.7826 (95% CI [0.7009, 0.8535], n=70) represents the statistical limitation of studying rare classes in the naturally occurring prevalence of ICDR. The EyePACS dataset is imbalanced by a factor of 95.9:1 between Grade 0 and Grade 4, i.e., more than 16x more imbalanced than the APTOS 2019 benchmark (~6:1). There are three priorities for extending the study in this direction: (i) validation of the model in independent PDR test sets from Messidor-2, APTOS 2019, and IDRiD, which provide PDR test data with different distributions; (ii) cost-sensitive losses (focal loss, class-balanced loss) and feature-space SMOTE for MTF features instead of image-level; (iii) prospectively collecting Grade 4 cases through clinic

collaborations, solving the fundamental data scarcity problem rather than an algorithmic one.

2. Evaluation only on a single dataset (EyePACS) restricts the evaluation of model performance over various acquisition protocols and demographic diversity, validation on Messidor-2, APTOS 2019, and IDRiD is the most important cross-dataset validation priority, complementary to the minority-class directions above.

3. Aggressive random projection (h=128, compression ratio 392:1) was empirically validated by F1-weighted=0.9701 but is theoretically suboptimal — $h \in \{512, 1024\}$ is recommended to improve distance preservation (JL ε: 0.81 → 0.40).

4. Comprehensive probability calibration analysis. While the chosen Random Forest classifier produces probabilistic outputs that are inherently better-calibrated than modern deep neural networks [22, 23] — a property reflected in our strong per-class AUC (0.9908–0.9982) — formal calibration metrics (Brier score, Expected Calibration Error, reliability diagrams) and post-hoc recalibration (Platt scaling, isotonic regression) are not reported in this submission. Comprehensive calibration analysis is elevated as the first priority of the external validation phase on Messidor-2, APTOS 2019, and IDRiD, given that cross-dataset calibration robustness — rather than single-dataset point estimates — is the regulatory-relevant criterion for clinical deployment [23].

   Furthermore, the framework does not currently include a dedicated out-of-distribution (OOD) detection mechanism for retinal pathologies unrelated to DR (e.g., isolated glaucoma, age-related macular degeneration, retinal vein occlusion). The random forest backbone itself implicitly has partial robustness properties due to the distributed probabilities on OOD samples [22] unlike the known overconfident behavior of CNN and transformer architectures [23]. The full-fledged OOD problem solving is incorporated into the second objective of the external validation phase with the following concrete action plan: (i) probability threshold triage of RF predict_proba outputs to identify unreliable predictions for review by an ophthalmologist (immediate, without retraining); (ii) cross pathology validation using public non-DR datasets (e.g., RIM-ONE for glaucoma, public AMD image datasets) to estimate the false positive rate on OOD data; (iii) incorporation of an explicit OOD detection module using feature space scoring (Mahalanobis distance or density based) as a preclassifier gate; (iv) multi-pathology approach expansion toward joint DR-AMD-glaucoma screening.

5. The provided standard deviations refer to the configuration-wise variability and not the training run stochasticity. As shown by Bouthillier et al. [24], considering the training run variability through multiple training runs with different random seeds and bootstrap confidence intervals is critical for proper comparison of machine learning benchmarks; this, together with the calibration evaluation stated in point (4), is considered to be a priority extension for the external validation phase on Messidor-2, APTOS 2019, and IDRiD.

6. Static-image scope. The framework was trained and evaluated exclusively on static fundus images, without longitudinal follow-up data on DR progression. MTF captures the structural correlates of DR severity grades as visible in single-visit fundus photographs, consistent with the static-feature definition of the ICDR

classification scale. The framework does not predict future DR progression, does not estimate patient-specific time-to-grade-change, and cannot validate inferred structural transitions against observed clinical trajectories. Longitudinal validation on follow-up cohorts (e.g., DRCR.net registry, multi-visit clinical datasets) is identified as a key future direction beyond the current static-image scope.

Future extensions include head-to-head comparison with transformer-based DR systems (Vision Transformers, Swin Transformers, hybrid CNN-transformer architectures) on matched cross-dataset evaluation protocols (Messidor-2, APTOS 2019, IDRiD). The calibration properties of transformer architectures follow the same overconfidence pattern documented for CNNs [23] and do not alter the Random Forest calibration advantage [22]; the comparison will therefore focus on accuracy-deployment trade-offs at matched data-protocol conditions.

6. Conclusion

The present study performs a systematic ablation study of adaptive preprocessing, morphological feature extraction, and lightweight classification for achieving efficient diabetic retinopathy (DR) classification, resulting in three main observations. Firstly, the variation of preprocessing parameters demonstrates a limited effect on DR classification: Five configurations of AMS-CLAHE with different image quality (SSIM +47.7%, PSNR +48.4%) showed a difference of only 0.41-percentage-points in F1-scores, questioning the necessity of preprocessing optimization as a crucial research direction. Secondly, MTF model learns structural pathophysiology not captured by traditional convolutional models: Explicit morphological modeling adds 1.89 percentage points of performance to the system. Lastly, the stability of classifiers should be considered equally important with peak accuracy in making decisions for deployment: Random Forest demonstrates high stability ($\sigma = 0.0002$) and a 0.9701 F1-score without hyperparameter tuning compared to less stable methods. Overall, the integrated RF-MTF system, which achieves an F1-weighted score of 0.9701 and F1-macro of 0.8910 with just 1–3 ms inference time, a 20–40 MB memory footprint, and 2-3 h of training time, proves that efficient five-class ICDR classification and clinical applicability can be achieved simultaneously through structural modeling, algorithmic stability, and computational efficiency. The minority class performance drop (Grade 4 F1 = 0.78, 95% CI [0.70, 0.85]) is clearly identified as the main direction for improvements. Four directions for future work before clinical application are suggested: External validation on Messidor-2, APTOS 2019, and IDRiD; targeted minority class performance improvement through cost-sensitive losses and SMOTE; extension of the random projection dimension ($h \in \{512, 1024\}$); and clinical pilots with validation by ophthalmologists.

References

[1] Teo, Zhen Ling, Yih-Chung Tham, Marco Yu, Miao Li Chee, Tyler Hyungtaek Rim, Ning Cheung, Mukharram M. Bikbov et al. "Global Prevalence of Diabetic Retinopathy and Projection of Burden Through 2045: Systematic Review and Meta-Analysis." Ophthalmology 128, no. 11 (2021): 1580-1591.

[2] Ting, Daniel Shu Wei, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah 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 318, no. 22 (2017): 2211-2223.

[3] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development And Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[4] Santos, Ana Rita, Luis Mendes, Maria Helena Madeira, Ines P. Marques, Diana Tavares, João Figueira, Conceição Lobo, and José Cunha-Vaz. "Microaneurysm Turnover in Mild Non-Proliferative Diabetic Retinopathy Is Associated with Progression and Development of Vision-Threatening Complications: A 5-Year Longitudinal Study." Journal of Clinical Medicine 10, no. 10 (2021): 2142.

[5] Wilkinson, Charles P., Frederick L. Ferris III, Ronald E. Klein, Paul P. Lee, Carl David Agardh, Matthew Davis, Diana Dills et al. "Proposed International Clinical Diabetic Retinopathy and Diabetic Macular Edema Disease Severity Scales." Ophthalmology 110, no. 9 (2003): 1677-1682.

[6] Zuiderveld, K. (1994). Contrast Limited Adaptive Histogram Equalization. In P. S. Heckbert (Ed.), Graphics Gems IV (pp. 474–485). Academic Press. https://doi.org/10.1016/B978-0-12-336156-1.50061-6

[7] Kirszenberg, Alexandre, Guillaume Tochon, Élodie Puybareau, and Jesus Angulo. "Going Beyond P-Convolutions to Learn Grayscale Morphological Operators." In International Conference on Discrete Geometry and Mathematical Morphology, Cham: Springer International Publishing, 2021, 470-482.

[8] Early Treatment Diabetic Retinopathy Study Research Group. "Grading Diabetic Retinopathy from Stereoscopic Color Fundus Photographs—An Extension of the Modified Airlie House Classification: ETDRS Report Number 10." Ophthalmology 98, no. 5 (1991): 786-806.

[9] Decencière, Etienne, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain et al. "Feedback on a Publicly Distributed Image Database: the Messidor Database." Image Analysis & Stereology (2014): 231-234.

[10] Breiman, L Random Forests. Machine Learning, 45(1), 5–32. 2001. https://doi.org/10.1023/A:1010933404324

[11] Parthasharathi, G. U., Vasantha Kumar, R. Premnivas, and K. Jasmine. "Diabetic Retinopathy Detection Using Machine Learning." Journal of Innovative Image Processing 4, no. 1 (2022): 26-33.

[12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785.

[13] Cortes, C., & Vapnik, V. Support-Vector Networks. Machine Learning, 20(3), 273–297. 1995. https://doi.org/10.1007/BF00994018

[14] Ul Haq, N., Waheed, T., Ishaq, K., Hassan, M. A., Safie, N., Elias, N. F., & Shoaib, M. (2024). Computationally Efficient Deep Learning Models for Diabetic Retinopathy Detection: A Systematic Literature Review. Artificial Intelligence Review, 57, Article 309. https://doi.org/10.1007/s10462-024-10942-9.

[15] Cuadros, Jorge, and George Bresnick. "EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening." Journal of diabetes science and technology 3, no. 3 (2009): 509-516.

[16] California Healthcare Foundation. (2015). Diabetic Retinopathy Detection [Data set]. Kaggle. https://www.kaggle.com/c/diabetic-retinopathy-detection

[17] Neogi, A. (2021). Diabetic Retinopathy Arranged [Data set]. Kaggle. https://www.kaggle.com/datasets/amanneo/diabetic-retinopathy-resized-arranged

[18] Cawley, Gavin C., and Nicola LC Talbot. "On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." The Journal of Machine Learning Research 11 (2010): 2079-2107.

[19] Dharrao, Deepak, Madhuri Dharrao, Shreyas Patil, Sangeeth Salvin, Prashant Ahire, and Yashwant Dongre. "AI-Driven Detection and Classification of Diabetic Retinopathy Stages Using EfficientNetB0." Discover Applied Sciences (2025).

[20] Arora, Lakshay, Sunil K. Singh, Sudhakar Kumar, Hardik Gupta, Wadee Alhalabi, Varsha Arya, Shavi Bansal, Kwok Tai Chui, and Brij B. Gupta. "Ensemble Deep Learning and EfficientNet for Accurate Diagnosis of Diabetic Retinopathy." Scientific Reports 14, no. 1 (2024): 30554.

[21] Nazir, Kinza, Jisoo Kim, and Yung-Cheol Byun. "Enhancing Early-Stage Diabetic Retinopathy Detection Using a Weighted Ensemble of Deep Neural Networks." IEEE Access 12 (2024): 113565-113579.

[22] Niculescu-Mizil, Alexandru, and Rich Caruana. "Predicting Good Probabilities with Supervised Learning." In Proceedings of the 22nd international conference on Machine learning, 625-632. 2005.

[23] Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." In International conference on machine learning, PMLR, 2017, 1321-1330.

[24] Bouthillier, Xavier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand et al. "Accounting for Variance in Machine Learning Benchmarks." Proceedings of machine learning and systems 3 (2021): 747-769.