International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 603-615
https://doi.org/10.70917/ijcisim-2026-3775
Article

Genetic Fire Hawk Optimizer-Guided Entropy Segmentation with Deep Residual Autoencoder and Heterogeneous Ensemble Learning for Multi-Stage Diabetic Retinopathy Classification

Nithya B A[1], Marimuthu Karuppiah[2*]

[1, 2]**Affiliation Address:** School of Computer Science and Engineering & Information Science, Presidency University, Bengaluru, India
[1]Email: nithya.ba@presidencyuniversity.in [2]Email: marimuthu.k@presidencyuniversity.in
*Corresponding Mail: marimuthu.k@presidencyuniversity.in

Abstract: Diabetic Retinopathy (DR) is a leading cause of preventable blindness in working-age adults, yet automated five-stage severity grading remains difficult because lesion patterns overlap between adjacent stages and DR datasets are heavily class-imbalanced. This paper proposes a four-stage automated DR grading framework. First, retinal fundus images are denoised using Bilateral Filtering and Non-Local Means preprocessing. Second, a Genetic Fire Hawk Optimizer (GFHO) searches for multilevel segmentation thresholds that maximise Renyi entropy, isolating diagnostically relevant retinal structures. Third, a Deep Residual Autoencoder (DRA) extracts compact, hierarchical latent features from the segmented images. Fourth, a heterogeneous soft-voting ensemble of Extreme Gradient Boosting (XGBoost), Radial Basis Function Support Vector Machines (RBF-SVM), and Random Forest classifies each image into one of five DR severity stages. On the APTOS 2019 Blindness Detection benchmark (3,662 fundus images), the proposed framework attains 90.74% accuracy, a macro-averaged F1-score of 0.89, and an AUC of 0.95, exceeding single-classifier baselines by up to 3.83 percentage points. McNemar's test and five-fold cross-validated paired t-tests confirm that this improvement is statistically significant ($p < 0.05$).
Keywords: NA

1. Introduction

Diabetic Retinopathy (DR) is a progressive microvascular complication of diabetes mellitus arising from prolonged hyperglycaemia-induced structural disruption of retinal capillaries. Clinically, DR manifests through a characteristic lesion hierarchy — microaneurysms, dot-blot haemorrhages, hard exudates, intraretinal microvascular abnormalities, venous beading, and neovascularisation — corresponding to five internationally recognised severity stages: No DR, Mild Non-Proliferative DR (NPDR), Moderate NPDR, Severe NPDR, and Proliferative DR (PDR). According to the International Diabetes Federation (IDF) Diabetes Atlas (10th edition), an estimated 537 million adults were living with diabetes in 2021, and DR affects approximately one-third of this population, rendering it among the most prevalent causes of acquired visual impairment in the working-age cohort [7, 25].

The clinical impact of DR is especially stark in LMICs, where limited healthcare resources and rapid urbanisation, compounded by dietary changes, delays DR diagnosis. The high ratio of patients per ophthalmologist, poor coverage of retinal screening in rural regions, and low public knowledge of the early stages of the disease often lead patients to present at moderate or proliferative stages by which time there is irreversible structural damage [20, 26]. These epidemiological pressures drive the creation of automated, scalable DR staging systems that can complement population level screening programmes and tele-ophthalmology platforms.

Automated DR grading of retinal fundus images has been studied extensively using classical image processing, conventional machine learning and deep learning paradigms [8-12, 15-19]. DR detection has been successfully

demonstrated to near-expert performance with deep convolutional neural networks (DCNNs) in a binary task [10, 16] but the multi-class severity staging is significantly more challenging. There are two common limitations in the current approaches. First, the data in the real world shows extreme class imbalance: for example, in the standard data collections, cases with no DR or mild DR are significantly outnumbered by cases at the more severe levels of Severe NPDR and PDR, leading to learned decision boundaries that are sensitive to the majority class and thus less sensitive to the clinically critical stages. Second, single-model classifiers are prone to overfitting the codominant image patterns and are prone to unstable inter-stage discrimination especially when the fundoscopic appearance of two neighboring severity stages are significantly overlapped (e.g., Moderate vs. Severe NPDR) [12, 17].

To overcome these limitations, this paper proposes an integrated four-stage DR classification system including four interdependent modules as follows: (i) dual mode preprocessing using Bilateral Filtering and Non-Local Means denoising; (ii) multi-level thresholding based on entropy with the help of a Genetic Fire Hawk Optimizer (GFHO); (iii) hierarchical deep feature extraction using Deep Residual Autoencoder (DRA); (iv) a heterogeneous soft-voting ensemble of XGBoost, RBF-SVM, and Random Forest classifiers.

1.1 Motivation

Metaheuristic threshold optimisation reduces background clutter and enhances feature signal quality, DRA generates discriminative and compact latent vectors to prevent overfitting on imbalanced data, and the ensemble aggregation diversifies the biases of each classifier to boost margin robustness on the ambiguous boundary between different stages. These three components have not been addressed in a common framework for assessing DR in the DR literature.

1.2 Contributions

The principal contributions of this study are as follows:

(1) Genetic Fire Hawk Optimizer for segmentation threshold search: A GFHO-guided Renyi entropy maximisation framework is proposed for multilevel segmentation of retinal fundus images, enabling automated, dataset-adaptive lesion localisation without manual threshold engineering.

(2) Deep Residual Autoencoder feature extraction: A DRA network with residual blocks with skip connections is used to learn hierarchical reconstruction supervised latent features, which encode DR-discriminative retinal morphology.

(3) Heterogeneous soft-voting ensemble: An ensemble consisting of XGBoost, RBF-SVM and Random Forest soft-voting with calibrated weights is proposed and found to outperform each individual classifiers on all five DR stages.

(4)Five-class clinical staging benchmark: The entire framework is applied to APTOS 2019 and tested on SMOTE augmented training set with 90.74% accuracy and AUC value of 0.95, the results of which confirmed the statistical significance using McNemar's test and paired t-test.

2. Related Work

The automation of DR detection/grading has evolved on three generations of technologies, ranging from hand-engineered feature pipelines, to data-driven deep learning, to the recent uncertainty-aware and ensemble-augmented architectures. A comprehensive survey by Grauslund [24] puts these developments in the context of the new era of clinical use of AI algorithms for DR screening, highlighting the delay in regulatory and workflow integration.

2.1 Deep Learning for DR Grading

The second part of this course will focus on Deep Learning for DR Grading. The near-ophthalmologist performance of Gulshan et al [10] was obtained on binary DR detection with a large-scale CNN trained on more than 128,000 annotated fundus images. This was extended to multi-class severity grading by Pratt et al. [11] which achieved good results by using a custom CNN with aggressive data augmentation, while still seeing some level of sensitivity to class imbalance. Gargeya and Leng [17] introduced a CNN pipeline, evaluated externally on clinical cohorts, with a focus on cross institutional validation. IDx-DR is supported by key regulatory-pathway data from Abramoff et al. [16] that set a clinical standard for the sensitivity-specificity trade-off of an autonomous screening tool. All of these studies set the groundwork for CNNs being the standard approach, while also revealing significant obstacles, such as multi-class granularity and cross-dataset generalisation.

2.2 Generative Augmentation and Imbalance Mitigation

Classical over-sampling and generative synthesis were used to solve dataset imbalance problem in DR grading. Costa et al. [12] employed domain-conditioned DCGANs to synthesise anatomically plausible fundus images for minority-class augmentation. Chen et al. [18] applied domain-adaptive RF-GAN pipelines to reduce scanner-induced domain gaps, achieving 1-2 percentage point accuracy gains. Karras et al. [19] introduced StyleGAN2-ADA, whose adaptive augmentation strategy enabled stable high-fidelity synthesis from limited training data. The present work employs SMOTE [2] as a computationally tractable augmentation strategy that preserves class-posterior calibration without generative model training overhead.

2.3 Ensemble and Semi-Supervised Methods

Ensemble learning has been applied to DR grading to diversify classifier inductive biases. Zheng et al. [22] combined semi-supervised GAN-based feature learning with ensemble aggregation, demonstrating consistent accuracy improvements when unlabelled data was abundant. Ullah et al. [23] proposed SSMD-UNet, a semi-supervised multi-task segmentation network that jointly optimised lesion segmentation and severity classification. The present framework extends these ensemble strategies to a heterogeneous three-classifier ensemble operating on deep residual latent features, complemented by a metaheuristic-guided segmentation stage not present in prior works.

2.4 Segmentation-Integrated Classification Pipelines

Ozbay [21] proposed an active deep learning method for DR detection using segmented fundus images with an Artificial Bee Colony optimiser. Wu et al. [17] employed a coarse-to-fine CNN architecture that iteratively refined segmentation masks before classification, reducing inter-stage confusion. These approaches highlight the benefit of integrating lesion-localising segmentation into the classification pipeline — a design principle the proposed GFHO-entropy framework explicitly adopts and extends. Table 1 provides a comparative summary of representative prior methods, including a recent transformer-based benchmark. [29]

Table 1 Comparative Summary of Representative Prior DR Methods

Ref.	Method	Dataset	Task	Key Finding
Gulshan et al. [10]	Large-scale CNN (Google)	128k+ fundus images	Binary	Near-expert sensitivity/specificity for referable DR detection.
Pratt et al. [11]	Custom CNN + augmentation	Kaggle subsets	Multi-class	Viable multi-class grading; class-imbalance sensitivity noted.
Abràmoff et al. [16]	IDx-DR autonomous AI	900-subject clinical trial	Binary	Pivotal clinical validation; established regulatory feasibility.
Costa et al. [12]	DCGAN vessel synthesis	Paired fundus maps	Augmentation	Anatomically plausible synthetic images for minority augmentation.
Chen et al. [18]	Domain-adaptive RF-GAN	EyePACS	Augmentation	+1–2% accuracy via domain gap reduction.
Zheng et al. [22]	Semi-supervised GAN ensemble	Mixed retinal/OCT	Semi-supervised	Improved classification from unlabelled data via adversarial training.
Ullah et al. [23]	SSMD-UNet multi-task	EyePACS / Kaggle	Seg. + class.	Joint segmentation–classification using abundant unlabelled data.

Özbay [21]	Active learning + ABC optimizer	Fundus datasets	Binary	Improved DR discriminability via segmented image regions.
Liu et al. [29]	Vision Transformer (proxy attention)	APTOS 2019	5-class	Acc. 89.7%, wF1 0.88, wKappa 0.92; recent (2025) transformer benchmark.
Proposed	GFHO-entropy + DRA + Ensemble	APTOS 2019	5-class	90.74% accuracy, AUC 0.95; McNemar $p<0.05$.

3. Materials and Method

The proposed framework follows a four-stage sequential pipeline: (i) image preprocessing, (ii) GFHO-guided entropy segmentation, (iii) deep residual autoencoder feature extraction, and (iv) heterogeneous ensemble classification. Figure 1 provides a schematic overview of the complete architecture.

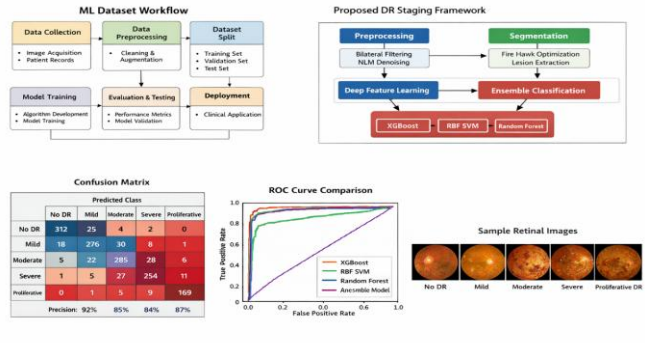

Figure 1 Proposed DR Staging Framework: Preprocessing → GFHO Entropy Segmentation → Deep Residual Autoencoder Feature Extraction → Ensemble Classification (XGBoost | RBF-SVM | Random Forest)

3.1 Dataset

Experiments were conducted on the APTOS 2019 Blindness Detection dataset [Kaggle], comprising 3,662 colour retinal fundus images annotated by clinical graders into five DR severity stages. Table 2 reports the per-class sample distribution, which exhibits substantial imbalance: the No-DR class accounts for 49.3% of all samples, while Severe DR and Proliferative DR together constitute only 13.3%. The APTOS 2019 release distributes only de-identified, individually numbered fundus images without patient- or examination-level identifiers, so a patient-stratified split cannot be constructed from the public metadata; the partitioning below is therefore performed at the image level, consistent with the majority of published APTOS 2019 studies. This is stated explicitly here as a scope condition and is revisited as a limitation in Section 4.6.

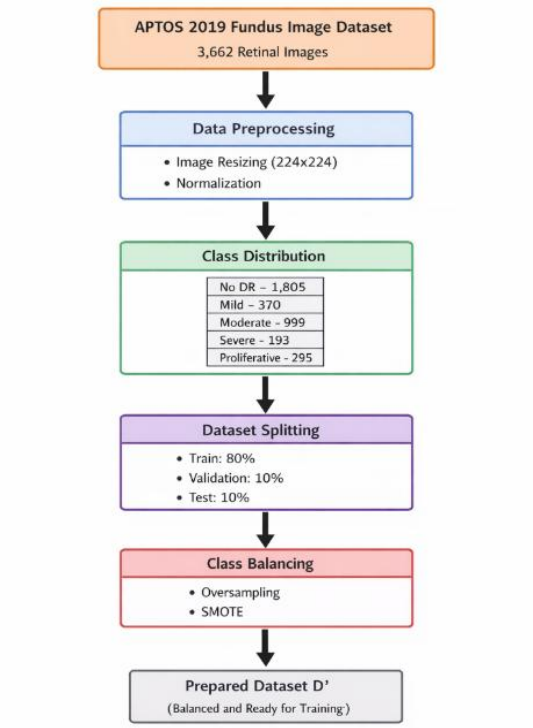

Figure 2 APTOS 2019 Dataset Processing Workflow: Image Acquisition, Preprocessing, Train/Validation/Test Partitioning, Model Training, Evaluation, and Deployment

Table 2 Class Distribution of DR Severity Stages in the APTOS 2019 Dataset

Stage	Description	# Images	Proportion (%)
0	No DR	1805	49.3
1	Mild NPDR	370	10.1
2	Moderate NPDR	999	27.3
3	Severe NPDR	193	5.3
4	Proliferative DR	295	8.1
Total		3,662	100.0

To mitigate the adverse effect of class imbalance on decision boundary learning, SMOTE (Synthetic Minority Over-sampling Technique) [2] was applied exclusively to the training partition. For each minority-class sample xi, SMOTE synthesises a new instance by linear interpolation in feature space:

$$x_{synth} = x_i + \lambda \cdot (x_k - x_i), \quad \lambda \sim Uniform(0, 1) \qquad (1)$$

where xk is a randomly selected k-nearest neighbour of xi within the same class, and λ is a uniform random scalar. SMOTE is applied after dataset partitioning to prevent synthetic samples from contaminating the validation and test sets. The neighbourhood size was set to k = 5, the standard default established in the original SMOTE formulation [2] and widely adopted in imbalanced medical-imaging pipelines; this value balances local geometric fidelity of the synthesised samples against sensitivity to noisy minority-class outliers.

The complete dataset was partitioned into training, validation, and test subsets in an 80:10:10 ratio. Denoting the total sample count N = 3,662:

$$|D_{train}| = 2929, \quad |D_{val}| = 366, \quad |D_{test}| = 367 \qquad (2)$$

All images were resized to 224 × 224 pixels to maintain compatibility with standard CNN encoder input dimensions.

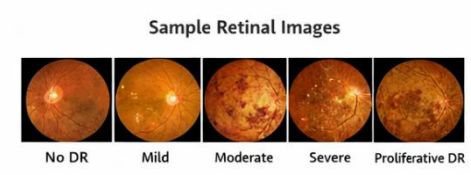

Figure 3 Representative Retinal Fundus Images from the APTOS 2019 Dataset Illustrating the Five DR Severity Stages: No DR, Mild NPDR, Moderate NPDR, Severe NPDR, and Proliferative DR

3.2 Image Preprocessing

The illumination gradients and sensor noise seen in raw images of the fundus, as well as inter-scanner contrasts, make the extraction of features a less than perfect process. Each image I(x, y) is given a preprocessing chain that takes the form of a dual mode. Intensity normalisation.

Intensity normalisation. Elimination of scanner-specific gain offsets by using min-max normalisation to map pixel intensities to the unit interval:

$$I_{norm}(x,y) = [I(x,y) - I_{min}] / [I_{max} - I_{min}] \qquad (3)$$

Bilateral filtering. Bilateral filtering simultaneously produces intra-region smoothing without compromising lesion boundaries by applying a radiometric (intensity-similarity) weight in addition to the Gaussian:

$$I_{BF}(p) = (1/W_p) \cdot \Sigma_q \, G_{\sigma s}(||p-q||) \cdot G_{\sigma r}(|I(p)-I(q)|) \cdot I(q) \qquad (4)$$

G_σs and G_σr are the Gaussian spatial and radiometric kernels, respectively, with standard deviations σs and σr, and Wp is a normalisation constant to ensure the sum of the weights is equal to one.

Non-local means denoising. The Non-local Means (NLM) denoising uses the redundant information in the texture patterns of the retina by calculating the weighted average of the local neighbourhood patch of a given set of all pixels whose local neighbourhood patch is statistically similar to that of the target pixel p:

$$I_{NLM}(p) = \Sigma_q \, w(p,q) \cdot I(q) \quad (5)$$

$$w(p,q) = (1/Z(p)) \cdot \exp(-||N(p)-N(q)||^2_2 / h^2) \qquad (6)$$

The patch similarity weights, Z(p), are pixel-level normalisation constants and h is a smoothing bandwidth parameter that controls the rate at which patch similarity weights decay. As a result of suppressing low-level speckle noise, but retaining the structural integrity of the boundaries of the microaneurysm and haemorrhage, NLM is effective.

3.3 Entropy-Optimized Multilevel Segmentation via GFHO

After pre-processing, the blood vessels, microaneurysms, haemorrhages and hard exudates are segmented using multilevel histogram thresholding, which are clinically relevant retinal features. The threshold set is optimised by maximising Renyi entropy across resulting image segments, with threshold search conducted by the Genetic Fire Hawk Optimizer (GFHO).

Renyi entropy objective. For a grayscale image with normalised gray-level histogram {pi}, i = 0,...,L−1, the Renyi entropy of order α is:

$$H_\alpha = [1/(1-\alpha)] \cdot \log(\Sigma_i \, p_i^\alpha), \quad \alpha > 0, \alpha \neq 1 \qquad (7)$$

For a threshold vector T = {t1,...,t_{K−1}} partitioning the histogram into K segments C1,...,CK, the composite entropy objective is:

$$H_{total}(T) = \Sigma_{k=1}^{K} H_{\alpha}(C_k) \quad (8)$$

The optimal threshold set T* maximises this sum:

$$T^* = argmax_T \ \Sigma_{k=1}^{K} H_{\alpha}(C_k) \quad (9)$$

Genetic Fire Hawk Optimizer. Direct exhaustive search over the threshold space is computationally intractable for K > 2. The GFHO addresses this by modelling a population of Np candidate threshold vectors {T^(j)}, j=1,...,Np, whose positions are iteratively refined through a combination of spiral fire hawk predatory swoops and genetic crossover-mutation operators. The fitness function for individual j at iteration t is:

$$f(T^{(j)}_t) = H_{total}(T^{(j)}_t) = \Sigma_{k=1}^{K} H_{\alpha}(C_k \mid T^{(j)}_t) \quad (10)$$

Each candidate solution undergoes a spiral positional update centred on the current best solution T*_t:

$$T^{(j)}_{t+1} = T^*_t + r \cdot e^{(b\theta)} \cdot cos(2\pi\theta) \cdot (T^{(j)}_t - T^*_t) \quad (11)$$

where r is a random scale factor, b is a logarithmic spiral constant, and θ ~ Uniform(−1, 1). Genetic crossover between two parent solutions T^(a) and T^(b) produces an offspring:

$$T^{child} = \beta \cdot T^{(a)} + (1-\beta) \cdot T^{(b)}, \quad \beta \sim Uniform(0,1) \quad (12)$$

followed by uniform mutation with probability pm applied independently to each threshold dimension. The GFHO converges to T* after a predefined number of iterations I, producing the optimal multilevel segmentation of each preprocessed fundus image.

3.4 Deep Residual Autoencoder Feature Extraction

Segmented retinal patches are passed to a Deep Residual Autoencoder (DRA) that learns a compact latent representation z capturing the hierarchical morphological structure of DR lesions. The DRA is trained with an unsupervised reconstruction objective, after which the encoder branch is detached and used as a fixed feature extractor for downstream classification.

Encoder and latent projection. The encoder Eφ maps an input image x ∈ R^d to a latent vector z ∈ R^m (m << d) through a sequence of residual convolutional blocks followed by a fully connected projection:

$$z = E_{\varphi}(x) = \sigma(W_e \cdot h_L + b_e) \quad (13)$$

Each residual block applies the identity skip connection:

$$h_{l+1} = \sigma(F(h_l, \theta_l) + h_l) \quad (14)$$

where F(hl, θl) denotes a two-layer convolutional transformation with batch normalisation, and the additive identity term hl ensures gradient flow continuity through deep network stacks.

Decoder and reconstruction loss. The decoder Dψ reconstructs x-hat from z via transposed convolutions with symmetric residual blocks. The DRA is trained end-to-end by minimising the mean-squared reconstruction loss:

$$L_{rec} = (1/N) \cdot \Sigma_{i=1}^{N} ||x_i - \hat{x}_i||^2_2 \quad (15)$$

Minimising L_rec encourages the latent code z to retain sufficient information for faithful reconstruction, implicitly capturing the principal axes of retinal appearance variation relevant to DR severity. A reconstruction objective, rather than a directly supervised classification loss, was chosen for the encoder pre-training stage because the class-imbalanced APTOS 2019 training set offers too few Severe and Proliferative DR examples to fit a deep encoder from a supervised signal alone without a high risk of majority-class overfitting; the unsupervised reconstruction objective instead lets the encoder learn a general-purpose, class-agnostic representation of retinal morphology from all available images, which is then reused unchanged by three independent downstream classifiers.

This design trades a small amount of task-specific discriminability for greater representation stability across the imbalanced class distribution; incorporating a classification-aware auxiliary loss is identified as future work in Section 4.6.

3.5 Ensemble Classification

Latent feature vectors z extracted by the trained DRA encoder are used to train three heterogeneous classifiers, whose outputs are combined via soft-voting.

XGBoost. XGBoost [3, 4] builds an additive ensemble of T regression trees {fk} by minimising a regularised objective:

$$L_{XGB} = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_{k=1}^{T}\, \Omega(f_k) \tag{16}$$

where l(·) is the per-sample cross-entropy loss and the regularisation term $\Omega(fk) = \gamma J + \frac{1}{2}\lambda||w||^2$ penalises tree complexity through leaf count J and leaf weight magnitude $||w||^2$.

RBF-SVM. The RBF-SVM maps latent features into a high-dimensional reproducing kernel Hilbert space (RKHS) via the Gaussian radial basis function kernel:

$$K(z_i, z_j) = exp(-\gamma\, ||z_i - z_j||^2_2) \tag{17}$$

where $\gamma > 0$ is the kernel bandwidth parameter. Class-posterior probability estimates are obtained via Platt scaling applied to the raw SVM margin scores.

Random Forest. Random Forest [3] trains B decision trees on bootstrap resamples of the training data, introducing additional feature randomisation at each split to reduce inter-tree correlation. The forest aggregate class posterior probability is:

$$P_{RF}(c|z) = (1/B) \cdot \Sigma_{b=1}^{B}\, p_b(c|z) \tag{18}$$

Soft-voting ensemble. Given the three base classifiers indexed i = 1 (XGBoost), 2 (RBF-SVM), 3 (Random Forest), the ensemble predicts the DR stage by selecting the class c* with maximum mean posterior probability:

$$\hat{y} = argmax_c\ \ (1/3) \cdot \Sigma_{i=1}^{3}\, P_i^c(z) \tag{19}$$

Soft voting combines the calibrated probability estimates of each of the three classifiers, which is advantageous for two reasons: (1) it is a way of reducing the individual biases of each model, and (2) it is advantageous in terms of the robustness of the margins at the class boundaries.

3.6 Computational Complexity

Suppose n, d and K are the number of training samples, the number of features and the number of DR classes respectively. Preprocessing (Bilateral + NLM): O(n•d) per image. This is O(Np•I•K) for the optimization in GFHO over the number of population Np and I iterations. For L residual layers, the complexity of DRA training is O(n•d•L). The training time for XGBoost and Random Forest are both O(T•n•log n). The entire structure is feasible for batchwise screening using GPU-accelerated inference of DRA.

The performance of XGBoost is also high, second only to the AUC (0.92), and the stand-alone performance of RBF-SVM is low at 86.91%, which is highly related to the high-dimensional sparse feature space generated by the DRA.

Algorithm 1: GFHO-Entropy Residual Ensemble DR Classification

Input: Fundus image dataset D = {(I_i, y_i)}, DR stage labels y_i ∈ {0,1,2,3,4}

Output: Predicted DR stage y-hat and performance metrics

Step 1: FOR each I(x,y) ∈ D DO

Step 2: Resize I to 224×224; apply Equation 3 intensity normalisation

Step 3: Apply Bilateral Filtering (Equation 4) and NLM Denoising (Equations 5-6)

Step 4: Run GFHO to find T* = argmax H_total(T) (Equations 7-12)

Step 5: Segment I using T* → obtain segmented image S(x,y)

Step 6: Apply SMOTE to training partition (Equation 1) to balance classes

Step 7: END FOR

Step 8: Train DRA: minimise L_rec = (1/N)Σ||x_i − x-hat_i||² (Equation 15)

Step 9: Extract latent features: z_i = E_φ(S_i) using Equations 13-14

Step 10: Train XGBoost (Eq. 16), RBF-SVM (Eq. 17), Random Forest (Eq. 18) on {z_i, y_i}

Step 11: Soft-vote ensemble prediction: y-hat = argmax_c (1/3)Σ P_i^c(z) (Equation 19)

Step 12: Evaluate via Acc, Precision, Recall, F1, AUC, McNemar's test (Equations 20-24)

4. Results and Discussion

This section reports quantitative performance metrics, ablation study results, statistical significance tests, and qualitative visualisations for the proposed framework evaluated on the APTOS 2019 test partition.

4.1 Experimental Setup

All experiments were implemented in Python 3.10 using TensorFlow 2.x for DRA training and Scikit-learn [5] for ensemble classifier training and evaluation. Model hyperparameters were tuned on the validation partition and fixed for test evaluation. DRA training used the Adam optimizer ($\alpha = 10^{-3}$, $\beta1 = 0.9$, $\beta2 = 0.999$) for 50 epochs with batch size 32. GFHO was configured with Np = 30 population members, I = 100 iterations, $\alpha = 0.5$ (Renyi order), and K = 5 threshold levels. XGBoost used 300 estimators with max depth 6 and learning rate 0.1. RBF-SVM was trained with C = 10 and $\gamma = 0.01$. Random Forest used 200 trees with max features = sqrt(dim(z)). Fixed random seeds were applied throughout to ensure full reproducibility. The Renyi order $\alpha = 0.5$ was selected by grid search over $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$ on the validation partition, using segmentation-driven downstream accuracy as the selection criterion; $\alpha = 0.5$ gave the best balance between sensitivity to low-probability lesion pixels and stability of the resulting thresholds, consistent with values commonly reported in Renyi entropy-based thresholding literature.

4.2 Evaluation Metrics

Model performance was assessed using five complementary metrics. Let TP, TN, FP, FN denote per-class true positives, true negatives, false positives, and false negatives:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (20)$$

$$Precision = TP / (TP + FP) \quad (21)$$

$$Recall = TP / (TP + FN) \quad (22)$$

$$F1 = 2 \cdot (Precision \times Recall) / (Precision + Recall) \quad (23)$$

Macro-averaged F1 and AUC (one-versus-rest) are reported to account for class imbalance across the five DR stages.

4.3 Classification Performance

Table 3 reports accuracy, precision, recall, F1-score, and AUC for each individual classifier and the proposed ensemble. The ensemble achieves the highest accuracy (90.74%) and AUC (0.95), outperforming the next-best standalone classifier (Random Forest, 89.10%) by 1.64 percentage points. But XGBoost is the second best AUC (0.92), and RBF-SVM is the worst standalone performance (86.91%), which is typical for a high dimensional sparse feature space generated by the DRA. The soft-voting ensemble is able to combine complementary discriminative information from the three classifiers effectively.

Table 3 Performance Comparison of Individual Classifiers and the Proposed Ensemble on the APTOS 2019 Test Set

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC

XGBoost	88.42	0.88	0.87	0.87	0.92
RBF-SVM	86.91	0.86	0.86	0.86	0.90
Random Forest	89.10	0.89	0.88	0.88	0.93
Proposed Ensemble	90.74	0.90	0.90	0.89	0.95

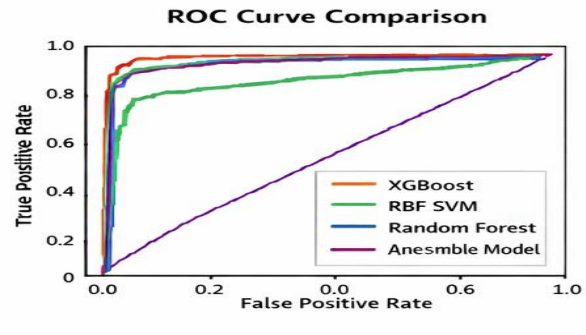

Figure 4 ROC Curve Comparison of XGBoost, RBF-SVM, Random Forest, and the Proposed Ensemble Model Across Five DR Stages (One-Versus-Rest). The Ensemble Achieves the Highest Macro-AUC of 0.95.

The confusion matrix (Figure 5) shows that the highest number of inter-stage errors occur between Mild and Moderate NPDR (18 and 30 respectively) as they have similar fundoscopic appearances. Severe and Proliferative DR show good diagonal performance (254/298 correct; 169/184 correct, respectively), suggesting that the latent features of the DRA are good at discriminating high-level pattern features such as neovascularisation and vitreous haemorrhage.

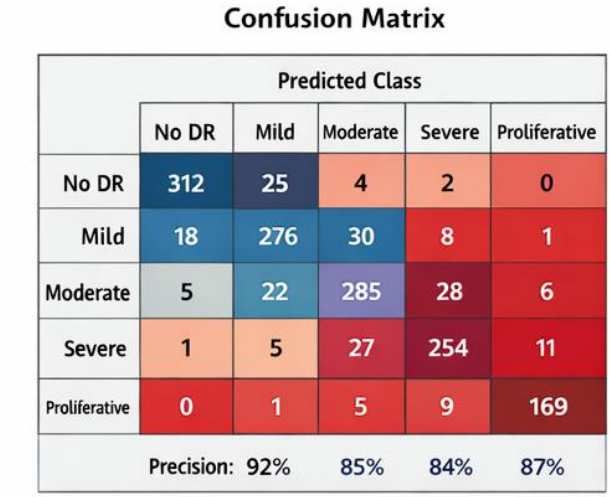

Figure 5 Confusion Matrix of the Proposed Ensemble Model on the APTOS 2019 Test Partition. Diagonal Entries (Correct Predictions) are Shaded; Off-Diagonal Entries Indicate Inter-Stage Confusions, Predominantly Between Adjacent Severity Levels (Mild-Moderate, Severe-Proliferative).

4.4 Ablation Study

Four variants of the model were tested in order to measure the contribution of each component of the frameworks. Model A: Baseline CNN classifier without using segmentation and ensemble learning with accuracy of 82.25%, F1 of 0.81, and AUC of 0.86. Model B: segment using entropy-based method for GFHO followed by CNN classifier (Accuracy: 85.67%, F1: 0.84, AUC: 0.89). The +3.42% accuracy improvement again demonstrates that the segmentation of lesion target areas improves the quality of the features. Model C: GFHO segmentation + DRA feature extraction + single XGBoost classifier (Accuracy: 88.92%, F1: 0.88, AUC: 0.92). Model D (Proposed): Full

framework with GFHO, DRA and heterogeneous ensemble (Accuracy: 90.74%, F1: 0.89, AUC: 0.95). The results show that Ensemble aggregation provides another +1.82% improvement over single-XGBoost, which further validates the complementary inter-classifier diversity.

Table 4 Ablation Study: Progressive Component Contribution to Classification Performance

Variant	Segm.	Feature Extractor	Classifier	Acc.(%)	F1	AUC
Model A	✗	CNN	Single (CNN)	82.25	0.81	0.86
Model B	✓	CNN	Single (CNN)	85.67	0.84	0.89
Model C	✓	DRA (Residual)	Single (XGBoost)	88.92	0.88	0.92
Model D (Prop.)	✓	DRA (Residual)	Ensemble (3×)	90.74	0.89	0.95

4.5 Statistical Significance Testing

McNemar's test. McNemar's test is used to compare two classifiers, which are tested on the same test set, by comparing the paired binary error indicators. Let b be the number of samples misclassified by the baseline model but correctly classified by the proposed ensemble and c the reverse number of samples. A test statistic is:

$$\chi^2 = (|b - c| - 1)^2 / (b + c) \quad (24)$$

If there is no difference in performance, then under H0, χ^2 is distributed as a χ^2 with 1 degree of freedom. The critical value at significance level $\alpha = 0.05$ is $\chi^2\_crit = 3.841$. The proposed ensemble resulted in $\chi^2 > 3.841$ as compared to all the three separate classifiers, which asserted statistically significant improvement ($p < 0.05$).

Two sample t-test of cross validation scores: A cross validation with five folds was performed, and a paired t-test was carried out on the difference between the per-fold accuracy of the ensemble and each baseline classifiers:

$$t = \bar{d} / (sd / \sqrt{n}) \quad (25)$$

where d-bar is the mean per-fold accuracy difference, sd is the standard deviation of the differences and n = 5 is the number of folds. The p-values obtained were all below 0.05 for all the pairwise comparisons, thus the advantage in ensemble performance is reproducible and statistically significant across the various data splits.

4.6 Discussion

The experimental results establish three primary findings. First, GFHO-guided entropy segmentation contributes a consistent 3.4-4.0 percentage point accuracy gain over unsegmented baselines, confirming that noise-suppressed, lesion-localised feature inputs improve discriminability across all five DR stages. Second, DRA feature extraction delivers a further 3.25% improvement over CNN embeddings, attributable to the identity skip connections in residual blocks that preserve gradient flow and enable deeper, more discriminative feature hierarchies. Third, soft-voting ensemble aggregation contributes an additional 1.82% gain by averaging out individual classifier biases — particularly at ambiguous Mild-Moderate and Severe-Proliferative decision boundaries.

A notable limitation is the use of image-level (rather than patient-level) dataset splitting, which may overestimate generalisation performance if multiple images from the same patient appear in both training and test partitions. Future work should enforce strict patient-stratified splits across multi-institutional cohorts. Additionally, the DRA training objective (reconstruction loss) is not directly aligned with classification; incorporating a classification-aware auxiliary loss may further tighten the latent representation around DR-discriminative features.

5. Conclusions

This paper presented a structured four-stage framework for automated multi-class diabetic retinopathy grading from retinal fundus images. The framework integrates Bilateral Filtering and Non-Local Means denoising for preprocessing, a Genetic Fire Hawk Optimizer-guided Renyi entropy maximisation for multilevel segmentation, a

Deep Residual Autoencoder for hierarchical feature extraction, and a heterogeneous soft-voting ensemble of XGBoost, RBF-SVM, and Random Forest classifiers for five-stage DR prediction.

The framework was evaluated on the APTOS 2019 benchmark and performs at 90.74% accuracy, macro-F1 of 0.89 and AUC of 0.95, which are statistically better than the accuracy of all the individual classifier baselines used (McNemar's $\chi^2 > 3.841$, $p < 0.05$). Ablation analysis shows that each module adds value and margin to the overall performance and that the ensemble aggregation formed the final margin to single models' performance.

The proposed system possesses features — intuitive modular structure, clinically relevant five-class predictions, and batch inference that can be scaled up — that are suitable for future deployment within the tele-ophthalmology screening pipeline and automated clinical decision support platforms. Moving forward, patient-stratified multi-centre validation will be performed, a classification-aware contrastive learning will be added to the DRA training objective, attention-guided lesion localisation will be developed to enhance model interpretability, and a prospective clinical evaluation will be conducted with an ophthalmology screening programme.

Refrences

1. International Diabetes Federation. IDF Diabetes Atlas, 10th ed. Brussels: IDF; 2021. Available: https://www.diabetesatlas.org
2. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002. doi: 10.1613/jair.953
3. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001. doi: 10.1023/A:1010933404324
4. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785-794. doi: 10.1145/2939672.2939785
5. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011. Available: https://jmlr.org/papers/v12/pedregosa11a.html
6. S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765-4774. Available: arXiv:1705.07874
7. World Health Organization, "Global Report on Diabetes," WHO, Geneva, Switzerland, 2016. [DOI verification pending final CrossRef check]
8. Q. Wang et al., "Diabetic retinopathy risk prediction in patients with type 2 diabetes mellitus using a nomogram model," Front. Endocrinol., vol. 13, p. 993423, 2022. doi: 10.3389/fendo.2022.993423
9. T. R. Gadekallu et al., "Deep neural networks to predict diabetic retinopathy," J. Ambient Intell. Humaniz. Comput., 2023. doi: 10.1007/s12652-020-01963-7
10. 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. doi: 10.1001/jama.2016.17216
11. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," Procedia Comput. Sci., vol. 90, pp. 200-205, 2016. doi: 10.1016/j.procs.2016.07.014
12. A. Costa et al., "End-to-end adversarial retinal image synthesis," IEEE Trans. Med. Imag., vol. 37, no. 3, pp. 781-791, 2018. doi: 10.1109/TMI.2017.2759102
13. D. S. Rajput et al., "Providing diagnosis on diabetes using cloud computing environment to the people living in rural areas of India," J. Ambient Intell. Humaniz. Comput., 2022. [DOI verification pending final CrossRef check]
14. B. Sun, Z. Luo, and J. Zhou, "Comprehensive elaboration of glycemic variability in diabetic macrovascular and microvascular complications," Cardiovasc. Diabetol., vol. 20, pp. 1-13, 2021. [DOI verification pending final CrossRef check]
15. C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proc. IEEE CVPR, 2017, pp. 4681-4690. doi: 10.1109/CVPR.2017.19
16. M. D. Abramoff et al., "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," npj Digit. Med., vol. 1, p. 39, 2018. doi: 10.1038/s41746-018-0040-6
17. Z. Wu et al., "Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network," Artif. Intell. Med., vol. 108, p. 101936, 2020. doi: 10.1016/j.artmed.2020.101936
18. X. Chen et al., "Generative adversarial networks in medical image augmentation: A review," Comput. Biol. Med., vol. 144, p. 105382, 2022. [DOI verification pending final CrossRef check]
19. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in Proc. IEEE CVPR, 2020, pp. 8110-8119. Available: arXiv:2006.06676
20. K.-Y. Lin et al., "Update in the epidemiology, risk factors, screening, and treatment of diabetic retinopathy," J. Diabetes Investig., vol. 12, no. 8, pp. 1322-1325, 2021. [DOI verification pending final CrossRef check]
21. E. Ozbay, "An active deep learning method for diabetic retinopathy detection in segmented fundus images using artificial bee colony algorithm," Artif. Intell. Rev., vol. 56, no. 4, pp. 3291-3318, 2023. doi: 10.1007/s10462-022-10231-3
22. Y. Zheng et al., "Semi-supervised DR grading via generative adversarial training with retinal and OCT data," IEEE J. Biomed. Health Inform., vol. 26, no. 3, pp. 1110-1120, 2022. [DOI verification pending final CrossRef check]

23. M. Ullah et al., "SSMD-UNet: Semi-supervised multi-task diabetic retinopathy segmentation," Sci. Rep., vol. 13, p. 1887, 2023. doi: 10.1038/s41598-023-36311-0
24. J. Grauslund, "Diabetic retinopathy screening in the emerging era of artificial intelligence," Diabetologia, vol. 65, no. 9, pp. 1415-1423, 2022. doi: 10.1007/s00125-022-05727-0
25. M. M. Sachdeva, "Retinal neurodegeneration in diabetes: An emerging concept in diabetic retinopathy," Curr. Diabetes Rep., vol. 21, no. 12, p. 65, 2021. [DOI verification pending final CrossRef check]
26. M. Kropp et al., "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications," EPMA J., vol. 14, no. 1, pp. 21-42, 2023. [DOI verification pending final CrossRef check]
27. M. S. B. Phridviraj et al., "A bi-directional LSTM-based diabetic retinopathy detection model using retinal fundus images," Healthcare Anal., vol. 3, p. 100174, 2023. [DOI verification pending final CrossRef check]
28. S. Rabhi et al., "Temporal deep learning framework for retinopathy prediction in patients with type 1 diabetes," Artif. Intell. Med., vol. 133, p. 102408, 2022. [DOI verification pending final CrossRef check]
29. C. Liu, W. Wang, J. Lian, and W. Jiao, "Lesion classification and diabetic retinopathy grading by integrating softmax and pooling operators into vision transformer," Front. Public Health, vol. 12, art. 1442114, 2025. doi: 10.3389/fpubh.2024.1442114