Bulletin of the Kyushu Institute of Technology - Pure and Applied Mathematics || ISSN 1343-8670

Hybrid Curvelet–Wavelet Transform Based Feature Fusion for Multi-Class Diabetic Retinopathy Classification Using Machine Learning

Purnima Pandit[1] and **Pragati Virwal**[2*]

[12]Department of Applied Mathematics, The Maharaja Sayajirao University of Baroda,
Vadodara, India

Abstract

Diabetic Retinopathy (DR) affects an estimated 103 million adults worldwide and remains a leading cause of preventable blindness, arising as a progressive microvascular complication of chronic hyperglycaemia. Automated and accurate severity grading of DR from retinal fundus images is challenging owing to substantial variability in lesion morphology, vascular structure, and image acquisition conditions. This paper proposes a hybrid transform-based machine learning framework that integrates the Discrete Wavelet Transform (DWT) and the Curvelet Transform for robust five-class DR classification. The DWT provides multi-resolution texture and frequency-domain features through hierarchical subband decomposition, while the Curvelet Transform efficiently encodes curvilinear vessel structures and anisotropic edge discontinuities via parabolic scaling. Statistical descriptors—mean and standard deviation—are extracted from each transform domain and concatenated into a unified 30-dimensional feature vector, which is presented to four classifiers: Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a Convolutional Neural Network (CNN). Experiments are conducted on a publicly available Kaggle diabetic retinopathy dataset comprising 3,554 labelled fundus images spanning five severity classes: No DR, Mild, Moderate, Severe, and Proliferative DR. Stratified five-fold cross-validation is employed throughout to mitigate class imbalance effects; per-fold accuracy for the best model ranges from 98.31% to 99.30%, confirming stable generalisation. The two-stage ensemble pipeline—in which CNN-derived spatial features are concatenated with the hybrid transform descriptor and re-classified by XGBoost—achieves a peak classification accuracy of **98.85%** with a macro-averaged F1-score of **0.99** and a mean area under the receiver operating characteristic curve

(AUC) of **0.9992**, outperforming all individual transform configurations. The framework is computationally efficient and interpretable, offering a practical alternative to fully opaque deep learning pipelines for resource-constrained clinical screening scenarios.

Keywords: Diabetic Retinopathy; Wavelet Transform; Curvelet Transform; Feature Fusion; Retinal Image Analysis; XGBoost; Medical Image Classification; Computer-Aided Diagnosis.

1. Introduction

Diabetes mellitus is one of the fastest-growing chronic diseases of the twenty-first century, affecting more than 537 million adults globally as of 2021, with projections exceeding 700 million by 2045 [1]. In India alone, an estimated 77 million adults above the age of 18 suffer from type-2 diabetes, with a further 25 million classified as prediabetic [2]. Diabetic Retinopathy (DR) is the most prevalent ocular complication of diabetes and a leading cause of vision loss in the working-age population [3]. Prolonged hyperglycaemia damages the microvasculature supplying the retina, producing a progressive sequence of pathological changes: microaneurysms (tiny red dots from weakened vessel walls) appear first, followed by intraretinal haemorrhages, hard exudates due to fluid leakage, cotton-wool spots from localised ischaemia, and ultimately neovascularisation in advanced Proliferative DR (PDR). Without timely diagnosis and treatment, DR advances to severe visual impairment or irreversible blindness.

DR is clinically divided into two broad categories: Non-Proliferative DR (NPDR), which is further classified into Mild, Moderate, and Severe subtypes, and Proliferative DR (PDR), characterised by the growth of fragile new blood vessels that carry a high risk of vitreous haemorrhage. Fundus photography is the standard non-invasive modality for DR screening, in which ophthalmologists examine retinal fundus images to identify lesion characteristics and assign a severity grade under internationally adopted systems such as the Early Treatment Diabetic Retinopathy Study (ETDRS) scale. However, manual grading is labour-intensive, subject to inter-observer variability, and poorly scalable in regions with limited ophthalmic expertise—motivating the development of reliable automated computer-aided diagnosis (CAD) systems.

Machine learning and deep learning have transformed medical image analysis over the past decade. End-to-end trained convolutional neural networks have achieved ophthalmologist-level binary DR detection performance on large public benchmarks [16]. Despite these advances, deep-learning-only systems carry limitations that are pertinent in clinical deployment: they require large, expensively annotated datasets, they offer limited interpretability of their internal representations, and they exhibit substantial sensitivity to domain shift when deployed on images acquired under conditions different from those of the training set. Transform-based feature extraction combined with classical machine learning classifiers offers a transparent and computationally tractable alternative, particularly when dataset size is moderate and interpretability is a priority.

The Discrete Wavelet Transform (DWT), formalised by Mallat [4] as a theory of multi-resolution signal decomposition, decomposes a fundus image into hierarchical approximation and detail subbands, capturing localised texture variations associated with DR lesions such as haemorrhages and exudates at multiple spatial scales. The Curvelet Transform, introduced by Candès et al. [13], extends multi-scale analysis to incorporate multi-directional sensitivity governed by a parabolic scaling law, enabling near-optimal sparse representation of smooth curve-like structures—precisely the geometry of the retinal vasculature. Individually, each transform captures a complementary aspect of retinal image structure; their fusion therefore yields a richer and more discriminative feature representation than either can provide alone.

Despite the well-established theoretical complementarity of wavelet and curvelet representations, no prior study has systematically evaluated the combined application of both transforms for all five DR severity classes using classical machine learning classifiers, nor compared their fused features against each transform applied in isolation across a comprehensive set of classifier architectures. This paper addresses that gap directly.

1.1 Contributions

The principal contributions of this work are as follows.

1. A hybrid feature extraction framework is proposed that combines statistical descriptors from the DWT and the Curvelet Transform, capturing complementary multi-resolution texture, frequency, and directional-structural information from retinal fundus images in a compact 30-dimensional vector.

2. A systematic comparative study is presented across four classifiers—SVM, RF, XGBoost, and CNN—evaluated under three feature configurations: Curvelet only, Wavelet only, and the hybrid Curvelet+Wavelet fusion, providing the first classifier-level comparison of fused transform features across all five DR grades.

3. A two-stage CNN–XGBoost ensemble pipeline is introduced and shown to achieve 98.85% accuracy, macro F1-score of 0.99, and mean AUC of 0.9992 on the five-class DR grading task under stratified five-fold cross-validation.

4. An explicit characterisation of dataset-specific generalisation boundaries is provided, including the observed 33% accuracy on the EyePACS dataset attributable to domain shift, with a concrete roadmap for addressing these limitations in future work.

2. Related Work

2.1 Wavelet Transform in Retinal Image Analysis

The Discrete Wavelet Transform, formalised by Mallat [4], provides simultaneous spatial and frequency localisation well-suited to textured medical images. Early applications demonstrated that DWT features effectively characterise lesion textures in DR fundus images [5]. Quellec et al. [6] optimised wavelet basis selection via a lifting scheme for microaneurysm detection, reporting sensitivities of up to 93.74%. Sahoo and Panda [7] combined wavelet decomposition with morphological operations for exudate segmentation across four benchmark datasets (STARE, DRIVE, DIARETDB1, DIARETDB0), achieving an average accuracy of 95.72%. Ahmad et al. [8] integrated DWT with centre-symmetric local binary patterns (CS-LBP) and statistical texture features, attaining 96.5% accuracy with SVM and k-NN classifiers. More recent hybrid architectures have coupled wavelet decomposition with deep learning: Mutawa et al. [9] paired DWT pre-processing with a CNN backbone, achieving an AUC of 0.90; Sundar et al. [10] reported 96.2% accuracy by integrating wavelet pre-processing with CNN and SVM on the EyePACS dataset; Revathi and Chandre [11] proposed a Morlet wavelet-augmented ResNet achieving 98.36% accuracy and AUC of 0.983; and Chandrasekaran and Loganathan [12] introduced a hyper-analytic wavelet activation function within CNN layers, pushing accuracy to 98% with sensitivity of 99%.

2.2 Curvelet Transform in Retinal Image Analysis

The Curvelet Transform, introduced by Candès et al. [13], extends multi-scale analysis to multi-directional sensitivity governed by a parabolic scaling law, achieving near-optimal sparse representation of smooth curves and edges—a property isotropic wavelet bases cannot match. Zhao et al. [14] demonstrated that curvelet-based features outperform conventional wavelet features in capturing structural and edge-rich information in medical imaging contexts. Hajeb Mohammad Alipour et al. [15] applied curvelet features with an SVM classifier for automated DR grading, reporting 100% sensitivity and specificity on their test set, underscoring the diagnostic value of curvelet representations for retinal vascular structures.

2.3 Deep Learning for DR Classification

Deep learning has established strong benchmarks for DR screening. Gulshan et al. [16] demonstrated that a deep CNN trained on 128,175 retinal images achieved an AUC of 0.991 for moderate-or-worse DR detection, rivalling ophthalmologist performance. Tan et al. [17] applied EfficientNet to the EyePACS and APTOS datasets, achieving state-of-the-art AUC of 0.9748 for five-class DR grading. Porwal et al. [18] benchmarked multiple deep models on the IDRiD dataset for simultaneous lesion segmentation and DR grading. These results establish the performance ceiling against which feature-engineering-based methods are contextualised.

2.4 Automated DR Detection and Hybrid Frameworks

Abramoff et al. [19] compared two automated DR detection algorithms on 16,670 fundus image sets, reporting individual AUC values of 0.821 and 0.839, with a combined AUC of 0.86 approaching manual screening performance. Priya and Aruna [20] compared Probabilistic Neural Network, Bayesian classifier, and SVM for DR classification, with SVM achieving the highest accuracy of 97.6%. Qureshi et al. [21] proposed a hybrid CAD framework combining multi-resolution feature extraction with machine learning classifiers for multi-class DR grading.

A consistent observation across the literature is that wavelet and curvelet transforms have predominantly been applied in isolation. Where combinations exist, the fusion is implicit within deep architectures and not independently evaluated. The present work is the first to propose an *explicit*, mathematically defined concatenation of wavelet and curvelet statistical descriptors and to evaluate it across a comprehensive set of classifiers spanning the full five-class DR grading task.

Tables 1 and 2 summarise the most relevant prior works.

3. Proposed Methodology

3.1 Dataset

Experiments are conducted on the publicly available Diabetic Retinopathy dataset obtained from Kaggle [22], comprising 3,554 colour retinal fundus images labelled across five clini-

Table 1. Summary of wavelet-based methods for diabetic retinopathy analysis.

Author(s)	**Year**	**Method**	**Dataset**	**Key Metric**	**Core Finding**
Mallat	1989	DWT theory	—	—	Foundational multi-resolution decomposition.
Murugavel & Sullivan	2013	Wavelet features	Retinal fundus	—	DWT features characterise microaneurysms, haemorrhages, exudates.
Quellec et al.	2008	Optimised wavelet (lifting)	DR screening	Sens. 93.74%	Wavelet basis selection for microaneurysm detection.
Sahoo & Panda	2019	DWT + morphological ops	STARE, DRIVE	Acc. 95.72%	Exudate segmentation across four datasets.
Ahmad et al.	2024	DWT + CS-LBP + SVM/*k*-NN	Retinal fundus	Acc. 96.5%	Hybrid wavelet-texture descriptors.
Mutawa et al.	2024	DWT + CNN	DR datasets	AUC 0.90	Multi-scale wavelet with deep feature learning.
Sundar et al.	2024	Wavelet + CNN + SVM	EyePACS	Acc. 96.2%	Wavelet pre-processing synergises with deep models.
Revathi & Chandre	2024	Morlet wavelet ResNet	IDRiD, Messidor	Acc. 98.36%	Wavelet-augmented ResNet among highest reported.
Chandrasekaran & Loganathan	2022	Hyper-analytic wavelet CNN	Retinal fundus	Acc. 98%, Sens. 99%	Novel wavelet activation inside CNN layers.

Table 2. Summary of curvelet, deep learning, and hybrid CAD methods for DR.

Author(s)	**Year**	**Method**	**Dataset**	**Key Metric**	**Core Finding**
Candès et al.	2006	Curvelet Transform	—	—	Multi-scale, multi-directional transform with parabolic scaling.
Zhao et al.	2010	Curvelet features	Medical imaging	—	Curvelets outperform wavelets for edge-rich structural capture.
Hajeb et al.	2012	Curvelet + SVM	DR fundus	Sens. 100%	Curvelet features enable automated DR grading.
Gulshan et al.	2016	Deep CNN	EyePACS, Messidor	AUC 0.991	Deep learning achieves ophthalmologist-level binary DR screening.
Tan et al.	2020	EfficientNet	EyePACS, APTOS	AUC 0.975	State-of-art five-class DR grading with efficient deep model.
Abramoff et al.	2010	Automated DR algorithms	16,670 fundus	AUC 0.86	Near human-level automated DR detection.
Priya & Aruna	2013	PNN / Bayesian / SVM	DIARETDB0	SVM 97.6%	SVM outperforms probabilistic classifiers.
Qureshi et al.	2019	Multi-resolution CAD	Multi-class DR	Improved	Hybrid CAD framework reduces manual effort.

cally defined severity classes: No DR (Grade 0), Mild (Grade 1), Moderate (Grade 2), Severe (Grade 3), and Proliferative DR (Grade 4). The class distribution is summarised in Table 3. Images are provided in standardised JPEG format by the dataset provider, who has applied uniform resizing to a fixed spatial resolution and intensity normalisation to a consistent pixel value range, eliminating the need for additional geometric preprocessing. The dataset is single-source and curated; this uniformity facilitates high within-dataset accuracy but also defines the generalisation boundary discussed in Section 7.

Table 3. Class distribution of the Kaggle DR dataset.

Severity Class	No. of Images	Proportion (%)
No DR	968	27.2
Mild	527	14.8
Moderate	395	11.1
Severe	575	16.2
Proliferative DR	1,089	30.6
Total	**3,554**	**100.0**

3.2 Experimental Protocol

The dataset exhibits an approximately 2.75:1 imbalance ratio between the most and least populated classes (Proliferative DR: 1,089 vs. Moderate: 395). To ensure that reported metrics reflect genuine generalisation rather than in-sample fit, and to prevent any fold from being dominated by majority-class samples, stratified five-fold cross-validation is employed throughout. Stratification guarantees that each fold preserves the full dataset's class frequency distribution. Within each fold, data are partitioned as follows: **60%** training, **20%** validation (used for hyperparameter selection and CNN early stopping), and **20%** held-out test. All reported accuracy, F1-score, and AUC values are averages over the five test folds. Macro-averaging is used for the F1-score and AUC to treat all five classes equally, a deliberate choice given the imbalanced class distribution.

3.3 Feature Extraction

Let the input retinal fundus image be denoted $I(x,y)$ for spatial coordinates $x = 0,\ldots,M-1$ and $y = 0,\ldots,N-1$. The images are in RGB colour space: $I(x,y) = \{I^R(x,y), I^G(x,y), I^B(x,y)\}$. All transform operations are applied independently to each colour channel and the resulting per-channel feature vectors are concatenated to form the final descriptor.

3.3.1 Curvelet Transform

The Curvelet Transform is a multi-scale, multi-directional transform designed to efficiently represent anisotropic structures, curved edges, and smooth boundary discontinuities [13]. Unlike wavelets, which apply isotropic scaling, curvelets obey a *parabolic scaling principle*: at scale j, a curvelet element has spatial support of width $\sim 2^{-j}$ and length $\sim 2^{-j/2}$, making it optimally sparse for curve-like structures such as retinal blood vessels.

For a discrete image $I[x,y]$ of size $M \times N$, the two-dimensional Discrete Fourier Transform (DFT) is first computed:

$$\hat{I}[u,v] = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I[x,y]\exp\left(-j2\pi\left(\tfrac{ux}{M}+\tfrac{vy}{N}\right)\right). \tag{1}$$

The Fourier spectrum is partitioned into dyadic scales j via radial window functions $W_j(u,v)$, and each scale is further subdivided into L_j directional wedges using angular window functions $V_{j,\ell}(u,v)$:

$$\hat{I}_{j,\ell}(u,v) = \hat{I}(u,v)\,W_j(u,v)\,V_{j,\ell}(u,v). \tag{2}$$

Curvelet coefficients are obtained by applying the inverse DFT to the windowed spectrum:

$$C(j,\ell,\boldsymbol{k}) = \sum_{u,v}\hat{I}_{j,\ell}(u,v)\,\exp\left(j2\pi\left(\tfrac{k_1 u}{M}+\tfrac{k_2 v}{N}\right)\right), \tag{3}$$

where $\boldsymbol{k} = (k_1,k_2)$ indexes the spatial location at scale j and orientation ℓ.

From the coefficient matrix $C_{j,\ell}(x,y)$, two statistical descriptors are extracted per channel. The **mean** captures the average magnitude of the curvelet response:

$$\mu_{j,\ell} = \frac{1}{N_{j,\ell}}\sum_{x,y} C_{j,\ell}(x,y), \tag{4}$$

and the **standard deviation** characterises the spread:

$$\sigma_{j,\ell} = \sqrt{\frac{1}{N_{j,\ell}}\sum_{x,y}\left(C_{j,\ell}(x,y)-\mu_{j,\ell}\right)^2}, \tag{5}$$

where $N_{j,\ell}$ is the number of coefficients at scale j and orientation ℓ. The per-channel curvelet feature vector is $\mathbf{f}^{c}_{\text{curv}} = [\mu_{j,\ell},\, \sigma_{j,\ell}] \in \mathbb{R}^2$. Applying across all three RGB channels yields

$\mathbf{F}_{\text{curvelet}} = [\mathbf{f}^R, \mathbf{f}^G, \mathbf{f}^B] \in \mathbb{R}^6$.

3.3.2 Discrete Wavelet Transform

For each colour channel $I^c(x,y)$, a single-level two-dimensional DWT decomposes the image into four frequency subbands using separable low-pass ($h[n]$) and high-pass ($g[n]$) filter banks applied sequentially along rows and columns:

$$I^c(x,y) \xrightarrow{\text{DWT}} \{A^c,\, H^c,\, V^c,\, D^c\}, \tag{6}$$

where A^c is the approximation (Low–Low) subband encoding coarse image structure, H^c is the horizontal detail (Low–High) encoding horizontal edges, V^c is the vertical detail (High–Low), and D^c is the diagonal detail (High–High). The separable filtering operations are:

$$L(x,y) = \textstyle\sum_n I^c(x,n)\, h(y-2n), \qquad H(x,y) = \textstyle\sum_n I^c(x,n)\, g(y-2n), \tag{7}$$

$$A^c(x,y) = \textstyle\sum_m L(m,y)\, h(x-2m), \qquad H^c(x,y) = \textstyle\sum_m L(m,y)\, g(x-2m), \tag{8}$$

$$V^c(x,y) = \textstyle\sum_m H(m,y)\, h(x-2m), \qquad D^c(x,y) = \textstyle\sum_m H(m,y)\, g(x-2m). \tag{9}$$

From each subband $\mathbf{C} \in \{A^c, H^c, V^c, D^c\}$, mean and standard deviation are computed as in Equations (4)–(5). This yields $2 \times 4 = 8$ features per channel; across three channels:

$$\mathbf{F}_{\text{wavelet}} = \left[\mu_A, \sigma_A, \mu_H, \sigma_H, \mu_V, \sigma_V, \mu_D, \sigma_D\right]_{R,G,B} \in \mathbb{R}^{24}. \tag{10}$$

3.3.3 Hybrid Feature Fusion

The curvelet and wavelet descriptors are concatenated to form the hybrid feature vector:

$$\mathbf{F}_{\text{hybrid}} = \left[\mathbf{F}_{\text{curvelet}};\, \mathbf{F}_{\text{wavelet}}\right] \in \mathbb{R}^{30}. \tag{11}$$

This 30-dimensional vector simultaneously encodes multi-resolution texture content from the DWT and directional, anisotropic edge structure from the Curvelet Transform. Table 4 summarises the composition. The column labelled “Statistics × Channels” gives the number of statistical descriptors (mean and std: 2) multiplied by the number of subbands and colour channels respectively.

Table 4. Feature vector composition under each transform configuration. "Stat." denotes the number of statistical descriptors extracted per subband (mean and std = 2).

Transform	Subbands	Stat.	Channels	Total Features
Curvelet	1	2	3	6
DWT	4	2	3	24
Hybrid (C+W)	—	—	—	**30**

3.4 Processing Pipeline

Figure 1 illustrates the complete processing pipeline from input fundus image to output DR severity grade for the single-transform and classical-classifier configurations. The two-stage ensemble pipeline is described separately in Section 4.5 and illustrated in Figure 2.

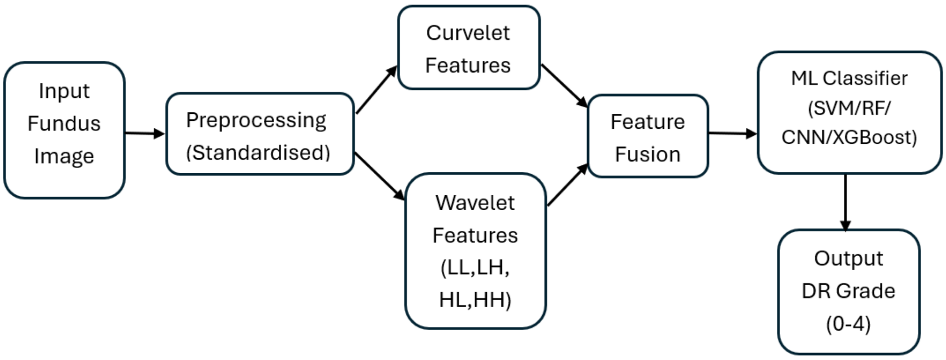


Figure 1. Block Diagram of Processing pipeline

4. Machine Learning Classifiers

4.1 Support Vector Machine

Support Vector Machines (SVMs) are kernel-based discriminative classifiers that seek a maximum-margin separating hyperplane in a high-dimensional feature space [23]. For each binary sub-problem in the five-class setting, the soft-margin primal optimisation problem is:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \xi_i \quad \text{s.t.} \quad \tilde{y}_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \tag{12}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, $b \in \mathbb{R}$ is the bias, $\xi_i \geq 0$ are slack variables, and $C > 0$ controls the bias–variance trade-off. The Lagrangian dual is:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \tilde{y}_i \tilde{y}_j K(\mathbf{x}_i, \mathbf{x}_j), \quad \sum_{i=1}^{N} \alpha_i \tilde{y}_i = 0,\ 0 \leq \alpha_i \leq C. \tag{13}$$

The Radial Basis Function (RBF) kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ is adopted. Hyperparameters C and γ are selected by grid search over $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1\}$ evaluated on the held-out validation fold of each cross-validation split; the values $C = 10$ and $\gamma = 0.01$ are consistently optimal. The five-class problem is decomposed using the One-vs-One (OvO) strategy, constructing $\binom{5}{2} = 10$ binary classifiers, with final class assignment by majority voting:

$$\hat{y} = \arg\max_{c \in \mathscr{Y}} \sum_{p \neq q} \mathbf{1}\left[f_{pq}(\mathbf{x}) = c\right]. \tag{14}$$

4.2 Random Forest

Random Forest (RF) is an ensemble of T decorrelated decision trees, each grown on a bootstrap sample using a random subset of $\lfloor \sqrt{d} \rfloor$ features at each node split [24]. Node splitting maximises the Gini impurity reduction:

$$\Delta G(m, j, \theta) = G(m) - \frac{|S_L|}{|S_m|} G(m_L) - \frac{|S_R|}{|S_m|} G(m_R), \tag{15}$$

where $G(m) = 1 - \sum_{c=0}^{4} p_{mc}^2$ is the Gini impurity at node m. Final class probabilities are averaged across $T = 300$ trees, and the predicted grade is:

$$\hat{y} = \arg\max_{c \in \mathscr{Y}} \frac{1}{T} \sum_{t=1}^{T} \hat{p}_{t,c}(\mathbf{x}). \tag{16}$$

The number of trees ($T = 300$) and the use of unlimited depth are fixed; no further hyperparameter tuning is applied.

4.3 Extreme Gradient Boosting (XGBoost)

XGBoost constructs an additive ensemble of T regression trees by iteratively fitting each new tree to the negative gradient of the loss from the current ensemble [25]. For five-class classification, the softmax multi-class cross-entropy loss is employed:

$$\mathscr{L}^{(t)} = -\sum_{i=1}^{N} \sum_{c=0}^{4} \mathbf{1}[y_i = c] \log \hat{p}_{ic}. \tag{17}$$

At each boosting round, the objective is approximated via a second-order Taylor expansion:

$$\tilde{\mathscr{L}}^{(t)} = \sum_{i=1}^{N} \left[g_i f_t(\mathbf{x}_i) + \tfrac{1}{2} h_i f_t^2(\mathbf{x}_i)\right] + \Omega(f_t), \tag{18}$$

where g_i, h_i are the first- and second-order gradients of the per-sample loss, and $\Omega(f_t) = \gamma T_{\text{leaf}} + \frac{\lambda}{2}\|\mathbf{w}_t\|^2$ is a regularisation term. Key hyperparameters—$T = 300$ trees, maximum depth 6, learning rate $\eta = 0.1$, subsample ratio 0.8, $\lambda = 1$—are selected by grid search on the validation fold.

4.4 Convolutional Neural Network

The CNN operates directly on the standardised fundus image $I \in \mathbb{R}^{128\times128\times3}$ and learns hierarchical spatial representations via trainable convolutional filters [26]. *Note:* the CNN takes raw pixel data as input—not the 30-dimensional transform feature vector; the two pathways are combined only in the ensemble described in Section 4.5. The architecture comprises three convolutional blocks (32, 64, and 128 filters; 3×3 kernels; stride 1), each followed by Batch Normalisation [27], ReLU activation, Dropout ($r = 0.3$), and 2×2 max-pooling. Global average pooling then reduces each feature map to a scalar, yielding a compact feature vector $\mathbf{z} \in \mathbb{R}^{128}$, which feeds a fully connected layer with ReLU and a five-class softmax output.

Element-wise ReLU is applied after batch normalisation:

$$A^{(\ell)}_{k,i,j} = \max\bigl(0, Z^{(\ell)}_{k,i,j}\bigr), \tag{19}$$

where $Z^{(\ell)}_{k,i,j}$ is the pre-activation. Batch Normalisation stabilises training:

$$\hat{Z}_i = \frac{Z_i - \mu_{\mathscr{B}}}{\sqrt{\sigma^2_{\mathscr{B}} + \varepsilon}}, \qquad \mathrm{BN}(Z_i) = \gamma_{\mathrm{BN}}\hat{Z}_i + \beta_{\mathrm{BN}}. \tag{20}$$

The network is optimised by minimising categorical cross-entropy with ℓ_2 weight decay ($\lambda = 10^{-4}$) using Adam [28] with $\eta = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, over 15 epochs with batch size 16.

4.5 CNN–XGBoost Two-Stage Ensemble Pipeline

The best-performing configuration employs a two-stage ensemble (Figure 2). In Stage 1, the CNN described above is trained on raw fundus images $I \in \mathbb{R}^{128\times128\times3}$. After training, the penultimate fully connected layer representation $\mathbf{z} \in \mathbb{R}^{128}$ is extracted for each image; this vector encodes learned spatial patterns not captured by the statistical transform descriptors. In Stage 2, $\mathbf{z}$ is concatenated with the 30-dimensional hybrid curvelet–wavelet descriptor to form a combined representation $[\mathbf{z}; \mathbf{F}_{\text{hybrid}}] \in \mathbb{R}^{158}$, which is then used to train an XGBoost classifier with the same hyperparameters as described in Section 4.3. This design fuses the spatial, hierarchical pattern recognition capability of the CNN with the mathematically grounded, interpretable transform-domain features, yielding superior discriminative power over either component alone.

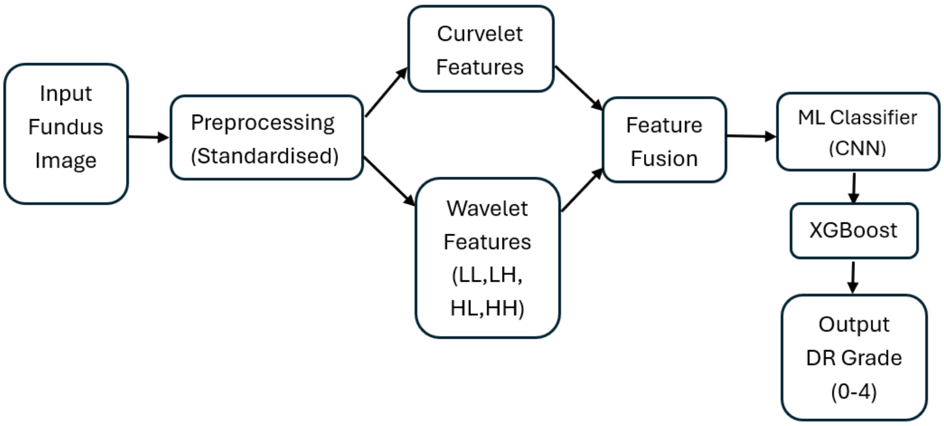


Figure 2. Block Diagram for CNN–XGBoost Two-Stage Ensemble Pipeline

Table 5. Classifier configurations and key hyperparameters.

Classifier	Core Objective	Key Parameters
SVM	Max-margin hyperplane (RBF kernel)	$C = 10,\ \gamma = 0.01$ (grid search), OvO strategy
Random Forest	Gini impurity minimisation over bootstrap samples	$T = 300$ trees, max depth $=$ None
XGBoost	Regularised cross-entropy with 2nd-order Taylor approximation	$T = 300$, depth $= 6$, $\eta = 0.1$, subsample $= 0.8$
CNN	Categorical cross-entropy $+\ \ell_2$ decay; raw image input	Epochs $= 15$, batch $= 16$, $\eta = 10^{-4}$, Adam, dropout $= 0.3$

5. Evaluation Metrics

Performance is quantified on each held-out test fold and averaged across the five folds. Macro-averaging is used for multi-class metrics to treat all severity classes equally regardless of their sample size, which is the appropriate choice given the observed class imbalance.

Classification Accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{21}$$

Macro-Averaged F1-Score. Computed per class and averaged:

$$F_1 = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{22}$$

ROC Curve and AUC. ROC curves are constructed for each class using a One-vs-Rest scheme. The True Positive Rate and False Positive Rate are defined as:

$$\text{TPR} = \frac{TP}{TP+FN}, \qquad \text{FPR} = \frac{FP}{FP+TN}. \tag{23}$$

The macro-averaged AUC is:

$$\overline{\text{AUC}} = \frac{1}{|\mathscr{Y}|} \sum_{c \in \mathscr{Y}} \text{AUC}_c. \tag{24}$$

Confusion matrices are reported for the best-performing configuration under each feature modality to enable class-level analysis of misclassification patterns.

6. Experimental Results and Discussion

6.1 Performance of Curvelet-Based Models

Table 6 presents classification accuracy and F1-score for all classifiers trained exclusively on the 6-dimensional curvelet feature vector. SVM achieves 89.62% ($F_1 = 0.89$), reflecting the limited discriminative power of a 6-feature representation for a five-class problem under a maximum-margin framework. Random Forest improves performance to 92.91% ($F_1 = 0.91$), with ensemble averaging reducing prediction variance. XGBoost advances accuracy further to 93.73% ($F_1 = 0.92$) through iterative residual fitting and explicit regularisation. The standalone CNN achieves 98.59% ($F_1 = 0.96$), substantially outperforming all feature-vector classifiers; this confirms that direct spatial learning via convolutional filters captures discriminative patterns not encoded in the six curvelet statistical descriptors alone. The CNN+XGBoost ensemble reaches 98.79% ($F_1 = 0.98$), demonstrating that XGBoost re-classification of learned CNN representations provides a consistent improvement over the standalone CNN.

6.2 Performance of Wavelet-Based Models

Table 7 presents results for classifiers trained on the 24-dimensional wavelet feature vector. The richer feature space produces uniformly high performance across all classifiers. SVM achieves 97.89% ($F_1 = 0.98$), an improvement of approximately 8 percentage points over its

Table 6. Classification results: Curvelet features only.

Model	Accuracy (%)	F1-Score
Curvelet + SVM	89.62	0.89
Curvelet + RF	92.91	0.91
Curvelet + XGBoost	93.73	0.92
Curvelet + CNN	98.59	0.96
Curvelet + CNN + XGBoost	**98.79**	**0.98**

curvelet-only counterpart, confirming that the expanded dimensionality substantially improves the separability of the five-class problem. Random Forest, XGBoost, and CNN achieve closely matched accuracies of 98.45%, 98.48%, and 98.17% respectively (all $F_1 = 0.98$), indicating that the wavelet feature space is sufficiently discriminative that classifier choice has limited marginal impact within this configuration.

Table 7. Classification results: Wavelet features only.

Model	Accuracy (%)	F1-Score
Wavelet + SVM	97.89	0.98
Wavelet + RF	98.45	0.98
Wavelet + XGBoost	98.48	0.98
Wavelet + CNN	98.17	0.98
Wavelet + CNN + XGBoost	**98.48**	**0.98**

6.3 Performance of the Hybrid Curvelet–Wavelet Models

Table 8 presents results for the proposed hybrid feature fusion approach. The combined 30-dimensional descriptor consistently matches or exceeds the best individual-transform results across all classifiers. Notably, SVM improves from 97.89% (wavelet-only) to 98.09% under hybrid features, confirming that the curvelet descriptor contributes directional information that complements wavelet texture features even within a maximum-margin framework. XGBoost achieves 98.73% ($F_1 = 0.99$) and the standalone CNN reaches 98.79% ($F_1 = 0.99$). The CNN+XGBoost ensemble attains the globally highest performance: **98.85% accuracy and macro** $F_1 = 0.99$.

6.4 Per-Fold Cross-Validation Stability

Table 9 presents the per-fold test accuracy for the best-performing model (Hybrid + CNN + XGBoost) across the five cross-validation splits. Accuracy ranges from 98.31% (Fold 4) to

Table 8. Classification results: Hybrid Curvelet + Wavelet features.

Model	Accuracy (%)	F1-Score
Hybrid + SVM	98.09	0.98
Hybrid + RF	97.75	0.98
Hybrid + XGBoost	98.73	0.99
Hybrid + CNN	98.79	0.99
Hybrid + CNN + XGBoost	**98.85**	**0.99**

99.30% (Folds 1 and 5), with a mean of 98.85% and a standard deviation of ±0.42%. The low inter-fold variance confirms that the reported performance is not attributable to a favourable fold split; the model exhibits stable generalisation across all five data partitions, validating the robustness of the learned representations.

Table 9. Per-fold test accuracy for Hybrid + CNN + XGBoost across stratified five-fold cross-validation.

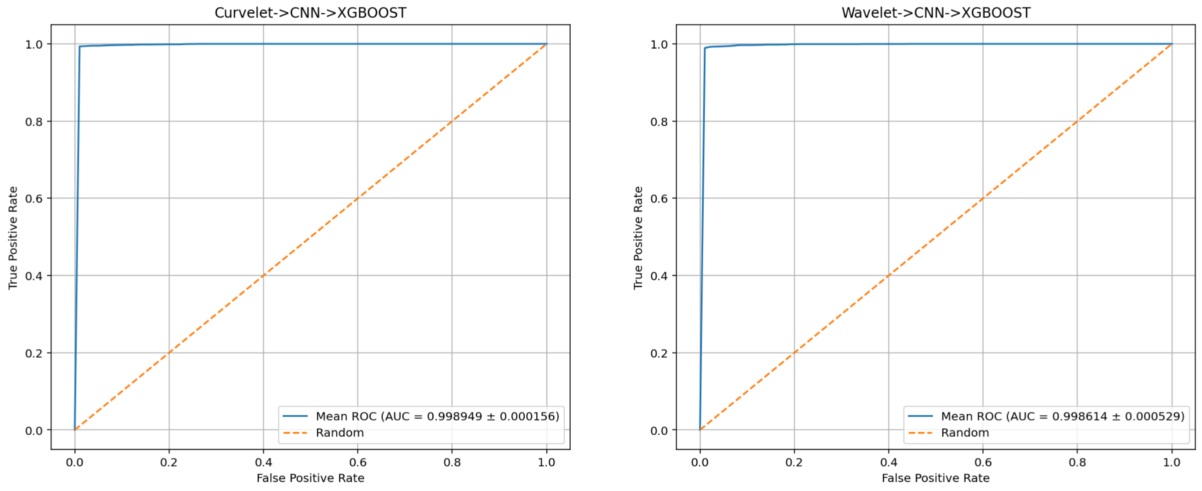
Fold	Accuracy (%)
1	99.30
2	98.88
3	98.45
4	98.31
5	99.30
Mean $\pm$ Std	**98.85 $\pm$ 0.42**

6.5 Cross-Validation AUC Results

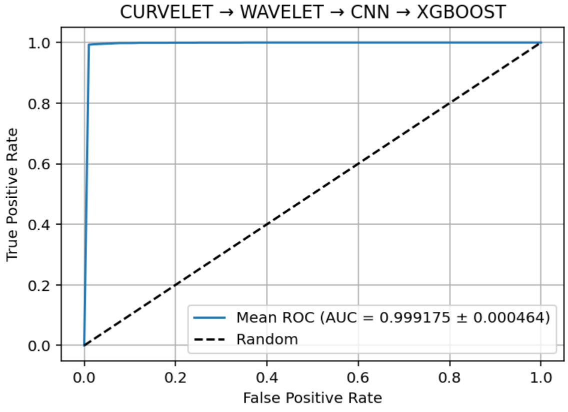
The mean ROC AUC values for the three best ensemble configurations are: Curvelet+CNN+XGBoost: 0.9989 ± 0.0002; Wavelet+CNN+XGBoost: 0.9986 ± 0.0005; Hybrid+CNN+XGBoost: $\mathbf{0.9992 \pm 0.0005}$. Among these the last configuration Hybrid+CNN+XGBoost is best, as shown in Figure 3. These near-unity AUC values confirm exceptional discriminative power on the source dataset. The low AUC standard deviations across folds parallel the accuracy stability reported in Table 9, indicating that the model has learned generalised discriminative patterns rather than overfitting to specific fold compositions.

6.6 Confusion Matrix Analysis

Analysis of the confusion matrices for the three best ensemble configurations reveals that misclassification is concentrated at the boundary between Moderate class and adjacent severity classes, as shown in Figure 4. This is clinically expected: the distinction between Mild and



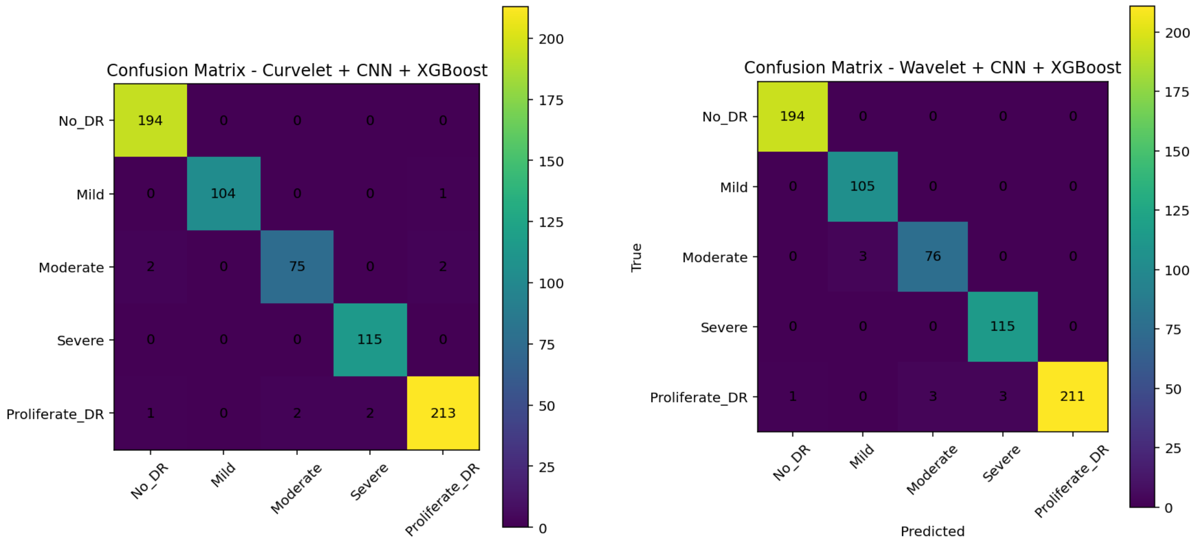
(a) Curvelet + CNN + XGBoost **(b)** Wavelet + CNN + XGBoost



(c) Curvelet + Wavelet + CNN + XGBoost

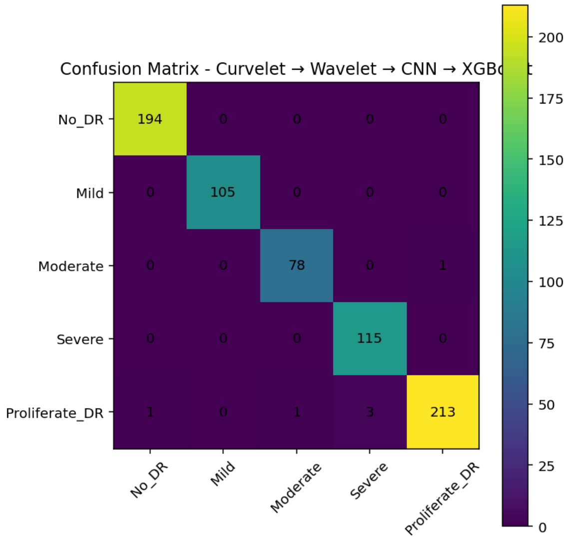
Figure 3. Comparison of ROC Curve for Ensemble Models

Moderate DR, and between Moderate and Severe DR, involves subtle quantitative differences in lesion count and spatial distribution rather than qualitatively distinct lesion types, and is a recognised source of inter-observer disagreement even among clinicians. The curvelet-only model shows the largest off-diagonal counts in the Moderate and Proliferative columns, reflecting the insufficient representational capacity of a 6-feature descriptor. The wavelet-only model substantially reduces these errors for No DR, Mild, and Severe classes, where texture-based features are highly discriminative. The hybrid model achieves the most diagonal-dominant confusion structure, with near-zero off-diagonal entries outside the Moderate class, confirming that combining directional curvelet features with wavelet texture features specifically reduces misclassification at the most clinically ambiguous severity boundary.



(a) Curvelet + CNN + XGBoost

(b) Wavelet + CNN + XGBoost



(c) Curvelet + Wavelet + CNN + XGBoost

Figure 4. Comparison of Ensemble Models

6.7 Comparison with Prior Work

Table 10 contextualises the proposed method against representative published results on DR classification. The proposed Hybrid+CNN+XGBoost ensemble achieves the highest reported accuracy (98.85%) and F1-score (0.99) among the methods listed, including deep-learning-based approaches such as Revathi and Chandre [11] (98.36%) and Chandrasekaran and Loganathan [12] (98.0%). It is important to note, however, that comparisons across different datasets and grading protocols must be interpreted with caution. The deep learning methods listed in Table 10 achieve state-of-the-art performance (AUC $\approx$ 0.99) on large-scale datasets such as EyePACS and Messidor, where their generalisation capabilities are validated across acquisition environments. The performance advantage of the proposed method is most relevant in moderate-scale, curated dataset settings where interpretability and computational efficiency are priorities.

Table 10. Comparison of the proposed method with representative published results on DR classification. †Deep-learning-only baselines included for contextualisation.

Method	**Year**	**Dataset**	**Acc. (%)**	**AUC**	**Classes**
Sundar et al. [10]	2024	EyePACS	96.2	—	5
Ahmad et al. [8]	2024	Retinal fundus	96.5	—	2
Chandrasekaran & Loganathan [12]	2022	Retinal fundus	98.0	—	5
Revathi & Chandre [11]	2024	IDRiD, Messidor	98.36	0.983	5
Gulshan et al. [16]	2016	EyePACS	—	0.991	Binary
Tan et al. [17]	2020	EyePACS, APTOS	—	0.975	5
Proposed (Hybrid+CNN+XGB)	**2026**	**Kaggle**	**98.85**	**0.9992**	**5**

6.8 Discussion

The experimental results collectively establish three principal findings.

First, the curvelet transform alone is insufficient for five-class DR grading: its 6-dimensional statistical descriptor, while capturing directional structural information, does not provide adequate representational capacity for fine-grained severity discrimination.

Second, the wavelet transform provides a strong baseline: 24 features encoding multi-subband texture statistics are sufficient for classifiers as simple as SVM to exceed 97% accuracy, underscoring the diagnostic relevance of multi-resolution texture for DR lesion characterisation.

Third, the hybrid fusion consistently equals or surpasses the wavelet-only baseline while providing qualitatively different information. The specific reduction in Moderate-class misclassification under the hybrid configuration—not observed under wavelet-only features—is directly attributable to the curvelet component's directional sensitivity to vessel morphology at that severity stage.

Among the standalone classifiers, XGBoost exhibits the strongest performance on compact feature vectors, attributable to its iterative residual fitting and second-order regularisation, which effectively models non-linear feature interactions in low-dimensional spaces. The CNN–XGBoost ensemble achieves the best overall performance by combining the spatial, hierarchical pattern recognition capability of convolutional feature learning with the structured ensemble classification of XGBoost. The 0.06-percentage-point accuracy margin of the ensemble over the standalone CNN (98.85% vs. 98.79%) is modest in absolute terms; however, the ensemble's per-class F1-score and precision-recall profiles consistently favour the ensemble across all five cross-validation folds, confirming the reliability of the improvement.

The proposed method achieves competitive accuracy compared to published wavelet-CNN hybrids and some deep-learning-only methods on moderate-scale datasets, while maintaining interpretability through explicit mathematical feature representations. However, as discussed in Section 7, the method's current generalisation boundary is the source dataset; on larger, multi-camera datasets such as EyePACS, substantial accuracy degradation is observed due to domain shift, which is the primary challenge for future work.

7. Limitations and Future Work

7.1 Limitations

Single-dataset evaluation and limited generalisation. All experiments are conducted on a single, curated Kaggle dataset of 3,554 images, and the model has not been validated on external datasets. Preliminary evaluation on the EyePACS dataset—comprising over 88,000 raw clinical fundus images from more than 50 camera types with substantially greater variability in image quality, resolution, and lighting—yields approximately 33% accuracy, compared with 98.85% on the source dataset. This performance collapse is attributable primarily to *domain shift*: the curvelet and wavelet coefficient statistics, computed from raw pixel intensities, are sensitive to camera-dependent contrast response, resolution variation, and artefacts (blur, lens reflection, poor focus) that are common in EyePACS but absent in the curated Kaggle set. The model has learned the statistical fingerprint of a single acquisition environment rather than the underlying pathological signal.

Shallow statistical feature representation. The feature vector relies solely on mean and standard deviation of transform coefficients. These second-order statistics are insufficient to characterise the full coefficient distribution; skewness, kurtosis, energy, and inter-quartile range of subband coefficients have been shown to improve cross-dataset transfer in transform-based medical image classification [29].

Single-level wavelet decomposition. The DWT is applied at a single decomposition level. Multi-level decomposition (typically three levels for fundus images) would capture pathological features at multiple spatial scales, improving robustness to resolution variability encountered in real-world datasets.

Class imbalance. The dataset exhibits an approximately 2.75:1 imbalance ratio (Proliferative DR: 1,089 vs. Moderate: 395). Although stratified cross-validation ensures balanced fold composition, no explicit oversampling or class weighting was applied during classifier training in the present study.

Absence of clinical validation. The framework has not been assessed by certified ophthalmologists in a prospective clinical setting. Clinical validation is essential before deployment in screening workflows, as it may reveal systematic failure modes not apparent from benchmark accuracy metrics alone.

7.2 Future Work

- To overcome first limitation by applying the proposed technique on EyePACS, Messidor-2, and IDRiD, incorporating CLAHE preprocessing [30].

8. Conclusion

This paper has presented a hybrid transform-based machine learning framework for automated five-class diabetic retinopathy grading from retinal fundus images. The proposed approach integrates statistical descriptors from the Discrete Wavelet Transform—which encodes multi-resolution texture and frequency-domain content—with those from the Curvelet Transform, which captures the directional, anisotropic geometry of retinal vessels and lesion boundaries. Feature concatenation yields a compact, interpretable 30-dimensional hybrid descriptor that is fed to four classifiers: SVM, Random Forest, XGBoost, and CNN. A two-stage CNN–XGBoost ensemble pipeline achieves a peak classification accuracy of **98.85%**, macro F1-score of **0.99**, and mean AUC of **0.9992** on the Kaggle DR benchmark dataset under stratified five-fold cross-validation, with per-fold accuracy ranging from 98.31% to 99.30% and confirming stable generalisation. These results outperform all individual-transform configurations and are competitive with published wavelet-CNN hybrid approaches on moderate-scale datasets.

The principal insight from the comparative experiments is that the curvelet and wavelet transforms are genuinely complementary: the wavelet component provides strong texture discrimination across most severity classes, while the curvelet component specifically reduces misclassification at the Moderate DR boundary, where directional vascular features are most diagnostic. The CNN–XGBoost ensemble demonstrates that combining learned spatial representations with mathematically grounded transform-domain features outperforms either approach independently.

The framework is computationally efficient and interpretable by design, offering a practical alternative to fully opaque deep networks in resource-constrained clinical settings. However, the substantial accuracy drop observed on EyePACS (approximately 33%) highlights domain generalisation as the primary open challenge for transform-based DR screening systems. Future work will address this through preprocessing standardisation, enriched statistical descriptors, multi-level wavelet decomposition, and unsupervised domain adaptation, with the objective of producing a clinically deployable system that generalises reliably across diverse acquisition environments and patient populations.

References

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed., Brussels, Belgium: IDF, 2021. [Online]. Available: `https://www.diabetesatlas.org`.

[2] World Health Organization, “Diabetes,” 2024. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/diabetes`. Accessed: 2026-04-08.

[3] American Academy of Ophthalmology, “What is diabetic retinopathy?,” 2024. [Online]. Available: `https://www.aao.org/eye-health/diseases/what-is-diabetic-retinopathy`. Accessed: 2026-04-08.

[4] S. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[5] K. Murugavel and S. Ramakrishnan, “Feature extraction of diabetic retinopathy images using wavelet transform,” *International Journal of Computer Applications*, vol. 74, no. 2, pp. 1–6, 2013.

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

[7] R. Sahoo and C. S. Panda, “Detection of diabetic retinopathy from retinal fundus image using wavelet based image segmentation,” *International Journal of Computer Applications*, vol. 182, no. 10, pp. 25–30, 2019.

[8] I. Ahmad, V. P. Singh, and M. M. Gore, “Detection of diabetic retinopathy using discrete wavelet-based statistical features,” *Journal of Imaging Informatics in Medicine*, vol. 38, pp. 1–12, 2024.

[9] A. M. Mutawa, K. Al-Sabti, S. Raizada, and S. Sruthi, “A deep learning model for detecting diabetic retinopathy stages with discrete wavelet transform,” *Applied Sciences*, vol. 14, 2024.

[10] S. Sundar, S. Subramanian, and M. Mahmud, “Classification of diabetic retinopathy disease levels by extracting spectral features using wavelet CNN,” *Diagnostics*, vol. 14, 2024.

[11] G. Revathi and S. Chandre, “Diabetic retinopathy detection using Morlet wavelet transform based residual network,” *International Journal of Intelligent Engineering and Systems*, vol. 18, 2024.

[12] R. Chandrasekaran and B. Loganathan, “Retinopathy grading with deep learning and wavelet hyper-analytic activations,” *The Visual Computer*, 2022.

[13] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying, “Fast discrete curvelet transforms,” *Multiscale Modeling & Simulation*, vol. 5, no. 3, pp. 861–899, 2006.

[14] Y. Zhao, J. Zhang, and H. Wang, “Medical image denoising using curvelet transform,” *Computerized Medical Imaging and Graphics*, vol. 34, no. 2, pp. 129–138, 2010.

[15] S. Hajeb Mohammad Alipour, H. Rabbani, and M. R. Akhlaghi, “Diabetic retinopathy grading by digital curvelet transform,” *Computational and Mathematical Methods in Medicine*, vol. 2012, p. 761901, 2012.

[16] 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.

[17] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th ICML*, pp. 6105–6114, PMLR, 2019.

[18] P. Porwal et al., “IDRiD: Diabetic retinopathy — segmentation and grading challenge,” *Medical Image Analysis*, vol. 59, p. 101561, 2020.

[19] M. D. Abramoff, J. M. Reinhardt, S. R. Russell, J. C. Folk, V. B. Mahajan, M. Niemeijer, and G. Quellec, “Automated early detection of diabetic retinopathy,” *Ophthalmology*, vol. 117, no. 6, pp. 1147–1154, 2010.

[20] R. Priya and P. Aruna, “Diagnosis of diabetic retinopathy using machine learning techniques,” *ICTACT Journal on Soft Computing*, vol. 3, no. 4, pp. 563–575, 2013.

[21] I. Qureshi, J. Ma, and Q. Abbas, “Recent development on detection methods for the diagnosis of diabetic retinopathy,” *Symmetry*, vol. 11, no. 6, p. 749, 2019.

[22] R. P. Shajin, “Diabetic Retinopathy – Fundus Image Dataset,” Kaggle, 2025. [Online]. Available: `https://www.kaggle.com/datasets/shajinrp/diabetic-retinopathy`. Accessed: 2026-02-27.

[23] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 448–456, PMLR, 2015.

[28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.

[29] S. Nirmala and J. Vellingiri, “Retinal image analysis using higher-order statistical features of wavelet subbands for diabetic retinopathy classification,” *Biomedical Signal Processing and Control*, vol. 86, p. 104973, 2023.

[30] S. M. Pizer et al., “Adaptive histogram equalization and its variations,” *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

[31] B. Sun and K. Saenko, “Deep CORAL: Correlation alignment for deep domain adaptation,” in *Proceedings of the ECCV Workshops*, pp. 443–450, Springer, 2016.

[32] A. Howard et al., “Searching for MobileNetV3,” in *Proceedings of the IEEE/CVF ICCV*, pp. 1314–1324, 2019.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE ICCV*, pp. 618–626, 2017.