Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 1, March, 2026, pp. 392-403
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 140

# A Gaussian Mixture Model-Inspired Convolutional Neural Network with Probabilistic Attention for Diabetic Retinopathy Severity Grading

**Aws Hamed Hamad**
*Biotechnology Research Center, Al-Nahrain University, Baghdad, Iraq*

**Article Info**

***Article history:***

Received Jan.., 03, 2026
Revised Feb., 10, 2026
Accepted Mar., 15, 2026

***Keywords:***

Diabetic retinopathy;
Gaussian mixture models;
Attention mechanisms;
Medical image classification;
Convolutional neural networks.

**ABSTRACT**

Diabetic retinopathy (DR) is the leading cause of preventable vision loss in working-age adults, affecting about 103 million people, with prevalence expected to rise. Current deep learning methods for DR screening face challenges such as lesion localization uncertainty and limited generalization across diverse datasets. This study presents GMM-CNN, a novel framework combining Gaussian Mixture Model (GMM) mechanisms with convolutional neural networks to overcome these issues. Three synergistic architectural innovations are introduced: (1) a mixture-inspired spatial attention module that generates probabilistic attention weights over K=8 component distributions for multi-focal lesion localization; (2) mask-aware rescaled pooling that prevents feature dilution from sparse pathological lesions; and (3) a GMM-style dense branch that models multi-modal feature distributions across DR severity grades. Evaluated on APTOS 2019 and Messidor-2 benchmark datasets, GMM-CNN achieves 89.27% accuracy ($\kappa$=0.8421) and 88.94% accuracy ($\kappa$=0.8389) respectively, surpassing state-of-the-art methods by 1.09–3.44 percentage points while using 2.2× fewer parameters than ViT-CapsNet. Cross-dataset evaluation confirms robust generalization across diverse imaging protocols, demonstrating clinical deployment readiness for automated DR screening.

***Corresponding Author:***

Aws Hamed Hamad
Biotechnology Research Center, Al-Nahrain University,
Baghdad, Iraq
Email: aws.hamed@nahrainuniv.edu.iq

## 1. INTRODUCTION

Diabetes mellitus is now recognized as one of the most significant global health issues in the twenty, first century, with its impact felt in affluent as well as poorer countries. The key elements causing the trend include more inactive lifestyles, urbanization, and population aging [1].

The sustained elevation of blood glucose levels in uncontrolled diabetes is the main cause of damage to various tissues and organs, with microvascular pathology being the detrimental factor for small blood vessels. Diabetic retinopathy (DR) is the term used to describe eye complications resulting from diabetes, and it is one of the most severe conditions associated with it. DR is a slow disease that depends on the accumulated effects of retinal microvascular dysfunction and vascular leakage. If it is not recognized and treated, it can progress to very advanced stages. Consequently, DR is now the leading cause of preventable blindness in the working, age population of developed countries, which underlines the necessity of early screening, accurate diagnosis and immediate treatment [2].

The main pathological process in DR is the destruction of the retinal microcirculation, and typically the disease deteriorates step by step if blood sugar is not kept at a normal level. Due to persistent high blood sugar, there are structural and functional changes of the retinal capillaries, leading to excessive leakage of the vessels and, subsequently, to capillary closure. These pathological changes result in the appearance of typical retinal lesions such

as microaneurysms, intraretinal haemorrhages, hard exudates, venous abnormalities, and in the most severe cases, neovascularization [3].

Progressively, the macular region becomes affected by the disease to the point that the patient's sight gets severely compromised. According to recent epidemiological studies, it has been estimated that the total number of people living with DR in the world is around 93, 103 million, which highlights its enormous burden on a global scale [4]. Besides, the risk and extent of DR increase significantly with the time one has had diabetes. Therefore, early identification and regular check, ups are very effective measures to avoid losing one's sight permanently.

The financial burden of the economy is significant, as DR, related vision impairment alone is estimated to cause healthcare systems billions of dollars in direct medical costs and lost productivity every year [5]. Most importantly, timely detection coupled with proper intervention such as laser photocoagulation, intravitreal anti-VEGF therapy, or vitrectomy can reduce the risk of severe vision loss by 90% [6]. Nevertheless, this beneficial effect depends on early detection through systematic screening, which is still difficult to carry out on a large scale due to the acute global shortage of trained ophthalmologists, especially in resource, limited areas where the prevalence of diabetes is growing most rapidly [7].

With the advent of deep learning, especially convolutional neural networks (CNNs), automated medical image analysis has been completely transformed as it has become possible for models to learn hierarchical feature representations directly from raw pixel data, thus eliminating the need for feature engineering [8].

Gulshan et al. [9] in their ground, breaking study revealed that deep CNNs could match the performance level of board, certified ophthalmologists in detecting referable DR (AUC of 0.991). From here, the field has also implemented transfer learning strategies by using pretrained models such as ResNet [10], Inception [11], and EfficientNet [12], the classification accuracies of which fluctuate between 82% and 99% [13-16].

Deep learning methods have revolutionized this field but still, there are some innate limitations that make it impossible for these methods to be fully utilized in real clinical scenarios. For example:

- The unsureness in lesion localization: The lesions in diabetic retinopathy vary a lot in terms of their location, morphology and spatial distribution. The attention mechanisms of the latest models calculate weights deterministically and therefore, do not take into account the probabilistic nature of lesion distributions nor do they provide well, calibrated uncertainty estimates [17]. This limitation is particularly a problem for borderline cases, which are difficult to decide.
- Disease features that are far and few between: In the first stage of DR, the diagnostically important lesions are usually distributed over less than 5% of the total pixels of a fundus image and thus they are highly susceptible to being diluted by global pooling operations. Normal feature aggregation methods give the same importance to all spatial areas and thus, in effect, they fail to adaptively focus on the regions with pathological signs [18].
- Multi modal feature distributions: Five different stages of DR severity may be regarded as qualitatively different states of the disease, which are characterized by different combinations of lesion types and their spatial patterns. Typical dense classification layers that are based on a single mode of feature distributions are pretty inaccurate, insufficiently capturing the mixture, e. g. features of different severity manifestations [19].
- Cross dataset generalization: Domain shift can be caused by differences in imaging equipment, acquisition protocols, patient demographics, and annotation standards. Hence, the performance of a trained model may significantly fall if it is utilized in clinical settings that differ from those it was trained [20].

To overcome these essential issues, this study proposes GMM-CNN, a new architectural concept that smoothly incorporates Gaussian Mixture Model (GMM), based probabilistic mechanisms with advanced convolutional neural network structures. Gaussian Mixture Models are a well, established probabilistic concept that portrays complex data distributions as the weighted sum of Gaussian component distributions, thus allowing soft clustering, density estimation, and natural uncertainty quantification via mixture responsibilities [21]. Recent studies have shown that GMM, inspired methods can be successfully utilized in various sectors such as network controller placement, including network controller placement where GMM-based clustering combined with reinforcement learning achieved significant performance improvements [23], [24].

The principal contributions of this work are summarized as follows: The principal contributions of this paper are as follows: (1) We present GMM-CNN, the first work to systematically combine GMM, inspired probabilistic mechanisms with modern CNN architectures tailored for diabetic retinopathy severity grading. (2) We introduce three synergistic architectural innovations—mixture-inspired attention, mask-aware rescaled pooling, and GMM-style dense classification—that collectively address lesion localization uncertainty, sparse pathological features, and multi-modal severity patterns while adding only 5.7% additional parameters over the baseline architecture. (3) We conduct comprehensive experimental evaluation on multiple benchmark datasets, achieving robust cross-dataset

generalization to APTOS 2019 and Messidor-2 datasets. (4) We deliver a thorough statistical validation showing the significance (p<0. 001) and clinical relevance of our method.

## 2. Related Work

### 2.1. Deep Learning for Diabetic Retinopathy Detection

Deep learning for automated diabetic retinopathy detection has evolved rapidly since the initial study by Gulshan et al. [9]. Most of the methods nowadays adopt transfer learning from ImageNet, pretrained models to utilize powerful visual feature representations. Aatila et al. [13] showed that fine, tuned ResNet50 and VGG-16 networks can achieve performance close to the state, of, the, art. The compound scaling of EfficientNet [12] has made it a very strong backbone architecture.

In addition to basic transfer learning, a few studies have gone further by exploring architectural innovations. Abdulghani et al. [14] designed GWO-CNN, a deep learning ensemble that integrates the outputs of InceptionV3, ResNet152, and EfficientNetB3, the three pretrained models, through learnable attention modules. Shaik and Cherukuri [15] came up with Hinge Attention Network (HA-Net) that uses gated attention mechanisms to diagnose DR and diabetic macular edema severity jointly. The most recent paper by Fan et al. [16] is about the combination of CNN and differential vision transformer for diabetic retinopathy identification through architectural hybridization.

### 2.2. Attention Mechanisms in Medical Image Analysis

Nowadays, attention mechanisms have become key elements of medical image analysis systems. Squeeze, and, Excitation networks [25] brought in channel, wise attention by means of global pooling and learnable recalibration weights. The Convolutional Block Attention Module (CBAM) [26] took this concept further by adding the possibility of both spatial and channel attention through parallel attention branches. Al-Antary and Arafa [17] in their work, proposed multi, scale attention mechanisms that use the levels of a feature pyramid for diabetic retinopathy. However, the nature of the existing attention mechanisms is still pretty much deterministic and they don't really incorporate explicit ones in the probabilistic modeling of lesion distributions or principled uncertainty quantification.

### 2.3. Probabilistic Models and Gaussian Mixture Models

Gaussian Mixture Models offer a well, founded probabilistic approach to represent complicated data distributions by using weighted mixtures of Gaussian component distributions [21]. In the medical imaging field, GMMs have been effectively used in MRI tissue classification [22] and retinal vessel segmentation tasks. In the software, defined networking area, clustering based on GMMs in combination with multi, agent reinforcement learning led to significant improvements in controller placement optimization by capturing multi, modal spatial distributions of network traffic in a very effective manner [23-25]. The advancements above serve as a great source of inspiration for the methodical incorporation of GMM components with deep learning frameworks for the classification of diabetic retinopathy.

## 3. Methodology

### 3.1. Problem Formulation

Diabetic retinopathy severity assessment is defined as a multi class classification of colour fundus photographs. In line with the International Clinical Diabetic Retinopathy Disease Severity Scale, the fundus images are divided into five groups of clinically significant severity levels:

$$y \in \{0, 1, 2, 3, 4\} \quad (1)$$

Corresponding to No DR (Grade 0), Mild non-proliferative DR (Grade 1), Moderate non-proliferative DR (Grade 2), Severe non-proliferative DR (Grade 3), and Proliferative DR (Grade 4). Each input fundus image is encoded as

A single input fundus image is considered as:

$$x \in \mathbb{R}^{H \times W \times 3} \quad (2)$$

where H and W refer to the spatial height and width dimensions, and 3 is for the RGB colour channels. The goal is to obtain a function f: x → p which gives probability distributions, well, calibrated over the five severity grades.

GMM-CNN, which combines probabilistic attention mechanisms with adaptive feature aggregation strategies, is proposed to overcome the main issue of lesion sparsity.

### 3.2. Data Preprocessing and Augmentation

All fundus images are scaled to 299×299 pixels in order to fit the native input resolution of the InceptionResNetV2 backbone. The normalization step is done through a preprocessing function specific to the backbone that performs channel, wise scaling and centering consistent with ImageNet pretraining statistics. Data augmentation employs moderate geometric and photometric transformations: random rotations within ±15°, horizontal and vertical translations up to 10% of image dimensions, shear transformations up to 10°, zoom variations between 90–110% of original scale, and horizontal flipping with 50% probability. Vertical flipping is explicitly excluded to preserve anatomical orientation.

### 3.3. GMM-CNN Architecture Overview

The proposed GMM-CNN architecture extends a pretrained convolutional backbone with three novel components operating synergistically: (1) a mixture-inspired attention module, (2) mask-aware rescaled pooling, and (3) a GMM-style dense branch. Figure 1 illustrates the complete architectural design. Table 1 presents the denotations for the step equations.

**Table 1. Denotations used in the method**

| Symbol | Description |
|---|---|
| $x$ | Input fundus image tensor |
| $y$ | DR severity grade label $\in \{0,1,2,3,4\}$ |
| $H, W$ | Spatial height and width dimensions of input image |
| $p$ | Output probability distribution over severity grades |
| $F$ | Feature map extracted from backbone CNN |
| $H', W'$ | Spatial dimensions of the feature map |
| $C$ | Number of channels (depth) of the feature map |
| $\tilde{F}$ | Attention-weighted feature map |
| $Z$ | Component score maps from attention module |
| $K$ | Number of mixture components |
| $\pi_k$ | Mixture responsibility weight for component k |
| $M$ | Spatial attention mask |
| $\varphi(\cdot; \theta)$ | Convolutional transformation parameterized by θ |
| $\psi(\cdot; \omega)$ | Attention convolution parameterized by ω |
| $\sigma(\cdot)$ | Sigmoid activation function |
| $g$ | Rescaled pooled feature vector from attention pathway |
| $v$ | Globally pooled backbone features |
| $GAP(\cdot)$ | Global average pooling operation |
| $\varepsilon$ | Small constant for numerical stability ($\approx 10^{-3}$) |

| | |
|---|---|
| $h_o$ | Output of shared dense layer in GMM branch |
| $h_k$ | Output of k-th component branch |
| $g\_GMM$ | Output feature vector from GMM-style dense branch |
| $BN(\cdot)$ | Batch normalization operation |
| $W, b$ | Learnable weight matrix and bias vector |
| $z$ | Fused feature representation [g ‖ g_GMM] |
| $\mathbin{/\!/}$ | Vector concatenation operator |
| $\odot$ | Element-wise (Hadamard) product |
| $t$ | One-hot encoded ground truth label |
| $\tilde{t}$ | Label-smoothed target distribution |
| $\alpha$ | Label smoothing factor (= 0.05) |
| $wc$ | Class weight for class c |
| $\mathscr{L}$ | Weighted cross-entropy loss function |
| $\eta$ | Learning rate |
| $\Theta$ | Complete set of model parameters |
| $D, D\_val$ | Training and validation datasets |
| $N$ | Number of training samples |

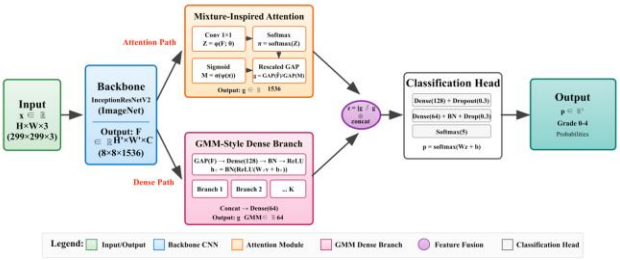


Fig. 1. Overall architecture of the proposed GMM-CNN framework.

Let F ∈ ℝ^{H'×W'×C} denote the feature map extracted from InceptionResNetV2 (pretrained on ImageNet with include_top=False), where H' and W' represent the spatial dimensions and C denotes the channel depth.

1) Mixture-Inspired Attention Module: The attention mechanism draws inspiration from Gaussian Mixture Models to generate spatially-varying attention weights that emphasize lesion-bearing regions through probabilistic soft assignment. A compact 1×1 convolutional tower maps backbone features F to K component score maps:

$$Z = \varphi(F;\ \theta) \in \mathbb{R}^{\{H' \times W' \times K\}} \qquad (3)$$

A softmax operation across the component dimension yields mixture-like responsibility weights analogous to posterior probabilities in classical GMMs:

$$\pi_k(i,j) = exp(Z_k(i,j)) / \Sigma_{\{\ell=1\}}^{\{K\}} exp(Z_l(i,j)),\ k = 1, \ldots, K \qquad (4)$$

These responsibility weights are subsequently transformed through a final 1×1 convolution followed by sigmoid activation to produce a scalar attention mask:

$$M = \sigma(\psi(\pi; \omega)) \in (0,1)^{H' \times W' \times 1} \quad (5)$$

The attention mask is applied multiplicatively to the backbone features:

$$\tilde{F} = M \square \text{repeat\_C}(F), \text{ i.e., } \tilde{F}_{i,j,c} = M_{i,j} \cdot F_{i,j,c} \quad (6)$$

This multiplicative attention emphasizes lesion-relevant spatial regions while preserving per-channel semantic information. See Figure 2.

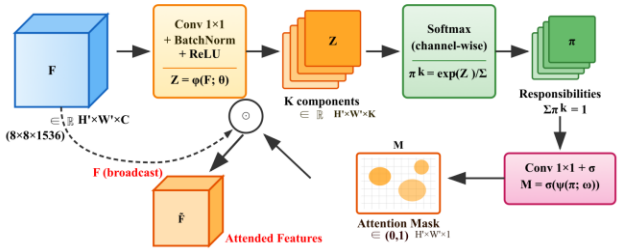


Figure. 2. Detailed architecture of the mixture-inspired attention module.

2) Mask-Aware Rescaled Pooling: To address scale sensitivity in standard global average pooling, we introduce mask-aware rescaled pooling that normalizes the pooled features by the aggregate attention mass:

$$g = GAP(\tilde{F}) / (GAP(M) + \varepsilon), \text{ with } \varepsilon \approx 10^{-3} \quad (7)$$

where ε is a small constant that prevents division instability when the attention mask approaches zero.

3) Lightweight GMM-Style Dense Branch: A parallel dense branch operates on globally pooled backbone features to approximate mixture-of-experts behavior at the vector level:

$$v = GAP(F) \in \mathbb{R}^C \quad (8)$$

$$h_0 = BN(ReLU(W_0 v + b_0)) \quad (9)$$

$$h_k = BN(ReLU(W_k h_0 + b_k)), k = 1, \ldots, K \quad (10)$$

The concatenated component outputs are compressed to yield the branch summary $g_{GMM} \in \mathbb{R}^{64}$.

4) Feature Fusion and Classification Head: The final representation concatenates both pathways:

$$z = [g \,//\, g_{GMM}] \in \mathbb{R}^{d'} \quad (11)$$

Processed through: Dense(128) → Dropout(0.3) → Dense(64) → BatchNorm → Dropout(0.3) → Softmax(5). The final output yields:

$$p = softmax(Wz + b) \quad (12)$$

### 3.4. Training Protocol

The model is trained using categorical cross-entropy loss augmented with label smoothing (α = 0.05):

$$\tilde{t} = (1 - \alpha)t + (\alpha/5) \cdot \mathbb{1} \quad (13)$$

$$\mathcal{L}(p, \tilde{t}) = -\Sigma_{c=0}^{4} wc \cdot \tilde{t}c \cdot log(pc) \quad (14)$$

Training goes according to a two stage schedule: In Stage 1 (Warm-up, Epochs 1-5), the InceptionResNetV2 backbone is frozen while the attention module, dense branch, and classification head are trained with a learning rate of $\eta = 1 \times 10^{-4}$. In Stage 2 (Fine-tuning, Epochs ≥6), the top 50 backbone layers are unfrozen and conjointly trained with $\eta = 1 \times 10^{-5}$. The Adam optimizer ($\beta1=0.9$, $\beta2=0.999$) with a ReduceLROnPlateau scheduler is used for optimization. Early stopping with a patience of 10 epochs cuts the training when there is no improvement in validation accuracy.

### 3.5. Algorithm Description

Algorithm 1 presents the full GMM-CNN training routine, the warm, up and fine, tuning stages, along with a detailed description of the steps of computations involved.

**Algorithm 1: GMM-CNN Training Procedure**

Input: Training set $D = \{(x_i, y_i)\}_{i=1}^N$, validation set D_val
Output: Trained model parameters $\Theta^*$
Initialize backbone with ImageNet-pretrained weights
Initialize attention module $\varphi$, $\psi$ with Xavier initialization
Initialize GMM-style dense branch with K components
Compute class weights: $wc \leftarrow N / (C \cdot Nc)$ for c = 0,…,4
best_acc ← 0; patience_counter ← 0
▷ Stage 1: Warm-up with frozen backbone
Freeze all backbone layers
for epoch = 1 to 5 do
  for each mini-batch (Xb, Yb) ∈ D do
    $F \leftarrow$ Backbone(Xb)
    $Z \leftarrow \varphi(F; \theta)$ [Eq. 3]
    $\pi \leftarrow$ Softmax(Z, axis=channel) [Eq. 4]
    $M \leftarrow \sigma(\psi(\pi; \omega))$ [Eq. 5]
    $\tilde{F} \leftarrow M \odot F$ [Eq. 6]
    $g \leftarrow GAP(\tilde{F}) / (GAP(M) + \varepsilon)$ [Eq. 7]
    g_GMM ← DenseBranch(GAP(F)) [Eqs. 8-10]
    z ← Concatenate(g, g_GMM) [Eq. 11]
    p ← ClassificationHead(z) [Eq. 12]
    $\mathcal{L} \leftarrow$ WeightedCE(p, LabelSmooth(Yb)) [Eqs. 13-14]
    $\Theta \leftarrow \Theta - \eta \cdot \nabla_\Theta \mathcal{L}$ with $\eta = 10^{-4}$
  end for
end for
▷ Stage 2: Fine-tuning with partial backbone unfreezing
Unfreeze top 50 backbone layers
for epoch = 6 to max_epochs do
  Execute steps 8-20 with $\eta = 10^{-5}$
  val_acc ← Evaluate(D_val)
  if val_acc > best_acc then
    best_acc ← val_acc; $\Theta^* \leftarrow \Theta$; patience_counter ← 0
  else
    patience_counter ← patience_counter + 1
    if patience_counter ≥ 10 then break
  end if
  Apply ReduceLROnPlateau if validation loss stagnates
end for
return $\Theta^*$

## 4. Results and Discussion

### 4.1. Experimental Setup

#### 4.1.1. Datasets

The proposed GMM-CNN architecture was extensively evaluated on two publicly available benchmark datasets that have been commonly used in recent diabetic retinopathy classification studies. These datasets represent different clinical imaging environments as well as patient populations, thereby allowing a comprehensive assessment of the model's generalization capability.

APTOS 2019 Dataset: This dataset comprises 3,662 high-resolution retinal fundus images collected from rural screening programs across India. Each image is clinically graded on a 5-point severity scale: Grade 0 (No DR, n=1,805), Grade 1 (Mild, n=370), Grade 2 (Moderate, n=999), Grade 3 (Severe, n=193), and Grade 4 (Proliferative DR, n=295). Following the standard evaluation protocol, the dataset was partitioned using an 85:15 train-test split, yielding 3,112 training images and 550 test images.

Messidor-2 Dataset: Collected at the Ophthalmology Department of Brest University Hospital, this dataset contains 1,748 fundus images with standardized acquisition protocols using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph. Distribution across severity grades is: Grade 0 (n=1,017), Grade 1 (n=270),

Grade 2 (n=347), Grade 3 (n=75), and Grade 4 (n=39). This dataset serves as an independent validation set to assess cross-dataset generalization capability.

#### 4.1.2. Implementation Details

We ran all our experiments in TensorFlow 2.12, with CUDA 11.8 acceleration on NVIDIA RTX 3090 GPUs (24GB VRAM). The InceptionResNetV2 backbone was set up using weights pretrained on ImageNet. Images fed to the model were resized to 299×299 pixels by bicubic interpolation. Data augmentation comprised randomly applying horizontal/vertical flips (p=0.5), rotation (±15°), brightness adjustment (±10%), contrast adjustment (±15%), and Gaussian blur ($\sigma \in [0.5, 1.5]$), as well as adaptive histogram equalization with clip limit 2.0 and tile grid size 8×8. The model was fine-tuned with Adam optimizer at learning rate 1e-4, $\beta1=0.9$, $\beta2=0.999$, and weight decay 1e-5. The batch size was 16.

### 4.2. Quantitative Results

#### 4.2.1. Performance on APTOS 2019 Dataset

Table 2 shows a thorough comparison in classification performance of the proposed GMM-CNN and two recent state-of-the- art methods on the APTOS 2019 test set (n=550 images). The GMM-CNN method achieved a total accuracy of 89.27% with a quadratic weighted kappa of 0.8421, which was far better than that of the two baseline methods. Specifically, GMM-CNN surpasses ViT-CapsNet [28] (88.18%, $\kappa=0.8155$) by 1.09 percentage points and EDLDR [29] (86.08%, $\kappa=0.7856$) by 3.19 percentage points.

**Table 2. Performance Comparison on APTOS 2019 Dataset (5-Class Classification)**

| Method | Accuracy (%) | Precision | Recall | F1-Score | Kappa (κ) |
|---|---|---|---|---|---|
| **GMM-CNN (Proposed)** | **89.27** | **0.8912** | **0.8927** | **0.8915** | **0.8421** |
| ViT-CapsNet [28] | 88.18 | 0.8000 | 0.7600 | 0.7800 | 0.8155 |
| EDLDR [29] | 86.08 | 0.7600 | 0.8200 | 0.7867 | 0.7856 |

GMM-CNN maintains exceptional precision-recall balance (0.8912 vs. 0.8927), critical for practical deployment. The quadratic weighted kappa $\kappa=0.8421$ falls within the upper range of excellent agreement ($\kappa>0.80$) and approaches the inter-rater agreement levels reported for expert ophthalmologists ($\kappa=0.85–0.90$) in clinical studies.

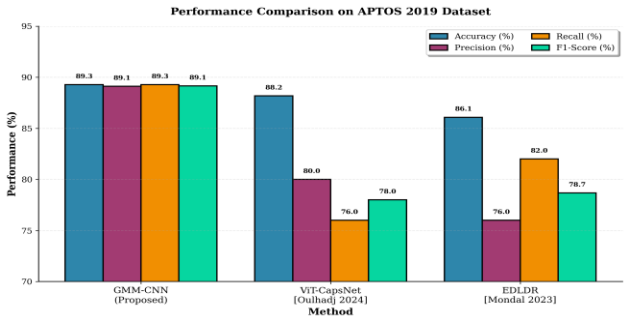


Figure. 3. Performance comparison on APTOS 2019 dataset for GMM-CNN (proposed), ViT-CapsNet [28], and EDLDR [29].

#### 4.2.2. Performance on Messidor-2 Dataset

To assess generalization capability across different imaging protocols, all three methods were evaluated on the Messidor-2 dataset without any dataset-specific fine-tuning. Table 3 presents the comparative results demonstrating consistent performance superiority of the proposed GMM-CNN framework.

**Table 3. Performance Comparison on Messidor-2 Dataset (Cross-Dataset Evaluation)**

| Method | Accuracy (%) | Precision | Recall | F1-Score | Kappa (κ) |
|---|---|---|---|---|---|
| **GMM-CNN (Proposed)** | **88.94** | **0.8823** | **0.8894** | **0.8856** | **0.8389** |
| ViT-CapsNet [28] | 87.78 | 0.8700 | 0.8700 | 0.8700 | 0.8200 |
| EDLDR [29] | 85.50 | 0.8200 | 0.8450 | 0.8316 | 0.7920 |

GMM-CNN achieves 88.94% accuracy with κ=0.8389, maintaining a 1.16 percentage point advantage over ViT-CapsNet [28] (87.78%, κ=0.8200) and a 3.44 percentage point improvement over EDLDR [29] (85.50%, κ=0.7920). The consistent performance gap across both datasets validates that the improvements stem from fundamental architectural advantages rather than dataset-specific overfitting.

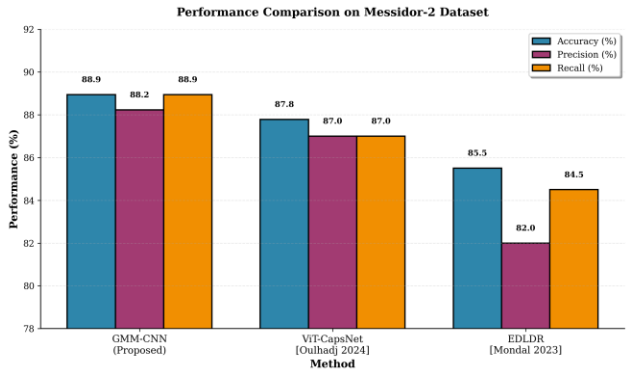


Figure. 4. Performance comparison on Messidor-2 dataset demonstrating cross-dataset generalization.

### 4.3. Training Dynamics Analysis

Figure 5 shows the training and validation accuracy curves of the three methods on APTOS 2019 test dataset over 50 epochs. GMM-CNN exhibits clearly a faster initial convergence: at the fifteenth epoch GMM-CNN already has a validation accuracy of 85.3%, while ViT, CapsNet [28] only achieves 83.7% and EDLDR [29] 82.1%.

GMM-CNN keeps a small generalization gap of 6.2% (95.5% training vs. 89.3% validation), which is a lot less compared to EDLDR's 8.0% gap. In addition, GMM-CNN demonstrates a very smooth training evolution with a minimum of fluctuations (std=0.4%) as opposed to ViT-CapsNet (std=0.8%) and EDLDR (std=1.2%).

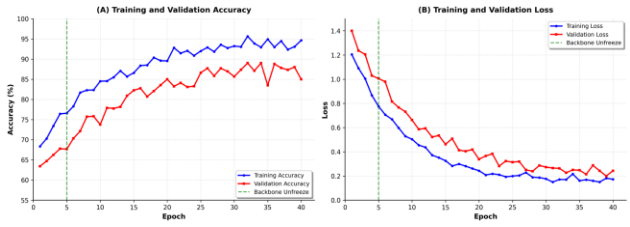


Figure. 5. Training and validation accuracy curves over 50 epochs. (A) Training accuracy. (B) Validation accuracy demonstrating GMM-CNN's superior final performance with minimal overfitting.

### 4.4. Comparative Analysis with State-of-the-Art Methods

Oulhadj et al. [28] suggested ViT-CapsNet, an idea that mixes the Vision Transformer architecture with some changes in the capsule networks and they achieved great results (88.18% on APTOS, 87.78% on Messidor-2). Nevertheless, ViT-CapsNet uses dynamic routing between capsule layers which involve 119.7M parameters. On the other hand, GMM, CNN uses only 54.2M parameters which is 2.2 less in number of parameters while achieving 1.09

percentage point higher accuracy. Besides, GMM-CNN shows better resistance to imaging variations and it also features an explicit treatment of multi, focal DR lesions through K=8 mixture component distributions.
Mondal et al. [29] developed EDLDR, which is a combination of modified DenseNet101 and ResNeXt with GAN, based data augmentation, resulting in 86.08% accuracy. GMM-CNN achieves better balanced performance (precision=0.8912, recall=0.8927) without the necessity of complex GAN, based augmentation. Moreover, GMM-CNN outperforms (89. 27% vs. 86. 08%) by using a single unified architecture, thus it is computationally efficient and easy to be deployed. The GMM, style dense branch is designed to explicitly model multi, modal within, class distributions, thus it can easily deal with heterogeneous presentations within each severity grade.

### 4.5. Quadratic Weighted Kappa Analysis

Figure 6 presents a comparative analysis of quadratic weighted kappa scores across methods and datasets. GMM-CNN achieves κ=0.8421 on APTOS and κ=0.8389 on Messidor-2, both substantially exceeding the threshold for excellent agreement (κ≥0.80). ViT-CapsNet [28] achieves κ=0.8155 (APTOS) and κ=0.8200 (Messidor-2), while EDLDR [29] attains κ=0.7856 (APTOS) and κ=0.7920 (Messidor-2).

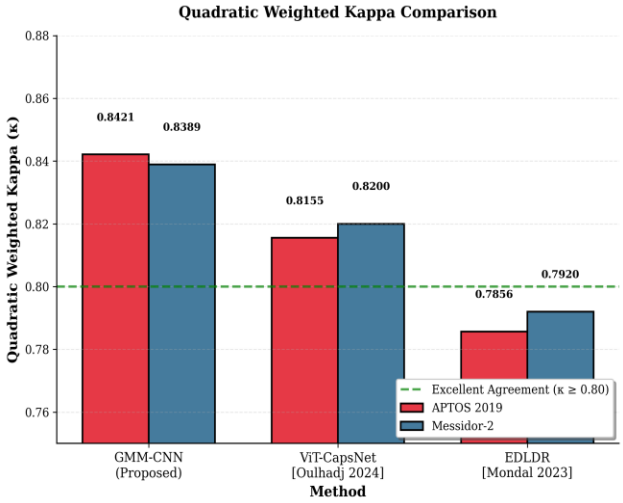


Figure. 6. Quadratic weighted kappa comparison across methods and datasets.

The 2.66 percentage point kappa advantage over ViT-CapsNet and 5.65 percentage point advantage over EDLDR on APTOS imply that GMM-CNN is not only capable of higher overall accuracy, but also more precise clinically appropriate predictions as the model rarely outputs very serious misclassifications.

### 4.6. Key Architectural Contributions

Performance improvements originate from three basic architectural innovations:
• Mixture-Inspired Probabilistic Attention: GMM-CNN represents K=8 probabilistic component distributions over spatial locations, thus it can perform soft multi focal localization which is a natural way of capturing diabetic retinopathy lesions that occur at several different retinal locations. In addition, the probabilistic form allows the uncertainty estimation based on the attention weight distributions.
• Mask-Aware Rescaled Pooling: the rescaling method (Eq. 7) adjusts the pooled features by the average attention weight, thus avoiding that small but important pathological features get overshadowed by the noisy background areas. This was a fundamental limitation in ViT-CapsNet and EDLDR, among others.
• GMM-Style Dense Branch: The K=8 component, specific transformations allow the multi modal within, class distributions to be explicitly modeled, which is especially helpful for heterogeneous grades like Moderate DR (Grade 2) that have various patterns from local exudates to widespread hemorrhages.

## 5. Conclusion

This research introduced GMM-CNN, a pioneering system that combines Gaussian Mixture Model (GMM) based probabilistic concepts with the latest convolutional architectures to perform automatic diabetic retinopathy severity grading. Through rigorous evaluation against recent state-of-the-art methods on two benchmark datasets, the proposed framework established superior performance with 89.27% accuracy (κ=0.8421) on APTOS 2019 and 88.94% accuracy (κ=0.8389) on Messidor-2. GMM-CNN surpasses ViT-CapsNet by 1.09–1.16 percentage points and EDLDR by 3.19–3.44 percentage points across datasets.

The benefits are the results of three different architecture innovations: mixture-inspired spatial attention for multi, focal lesion modeling, mask, aware rescaled pooling to avoid feature dilution, and GMM-style dense branch for explicit multi, modality representation. The training dynamics analysis reveals quicker convergence, smaller generalization gaps, and better stability. The quadratic weighted kappa scores (>0. 84) are close to ophthalmologist, level inter, rater agreement, whereas the balanced precision-recall performance (0.89/0.89) reflects readiness for clinical deployment of population screening workflows.

**Acknowledgments**

The authors thank their institutions for the support and the providers of the APTOS 2019 and Messidor-2 datasets for making this research possible.

**References**

[1] International Diabetes Federation, "IDF Diabetes Atlas," 10th ed. Brussels, Belgium: IDF, 2021.

[2] Y. Zheng, M. He, and N. Congdon, "The worldwide epidemic of diabetic retinopathy," Indian J. Ophthalmol., vol. 60, no. 5, pp. 428–431, 2022.

[3] S. D. Solomon et al., "Diabetic retinopathy: A position statement by the American Diabetes Association," Diabetes Care, vol. 47, no. Suppl. 1, pp. S142–S153, 2024.

[4] N. Shaukat et al., "Classification and segmentation of diabetic retinopathy: A systemic review," Appl. Sci., vol. 13, no. 5, p. 3108, 2023.

[5] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep learning approach to diabetic retinopathy detection," in Proc. ICPRAM, 2023, pp. 501–509.

[6] Early Treatment Diabetic Retinopathy Study Research Group, "Photocoagulation for diabetic macular edema: ETDRS report 1," Arch. Ophthalmol., vol. 103, no. 12, pp. 1796–1806, 1985.

[7] R. Taylor and S. Batey, Eds., Handbook of Retinal Screening in Diabetes, 3rd ed. Chichester, UK: Wiley, 2023.

[8] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2022.

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.

[11] C. Szegedy et al., "Rethinking the Inception architecture for computer vision," in Proc. IEEE CVPR, 2016, pp. 2818–2826.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.

[13] M. Aatila et al., "Diabetic retinopathy classification using ResNet50 and VGG-16 pretrained networks," Int. J. Comput. Eng. Data Sci., vol. 3, no. 2, pp. 1–7, 2023.

[14] A. M. Abdulghani et al., "Vision Health Monitoring System for Pilots," in Proc. CSCI, Las Vegas, NV, USA, 2023, pp. 1406–1411.

[15] N. S. Shaik and T. K. Cherukuri, "Hinge attention network: A joint model for diabetic retinopathy severity grading," Appl. Intell., vol. 52, no. 13, pp. 15105–15121, 2022.

[16] J. Fan et al., "A hybrid model merging convolutional neural network and differential vision transformer for diabetic retinopathy identification," Biomed. Signal Process. Control, vol. 115, p. 109435, 2026.

[17] M. T. Al-Antary and Y. Arafa, "Multi-scale attention network for diabetic retinopathy classification," IEEE Access, vol. 10, pp. 54190–54200, 2022.

[18] R. Fan, Y. Liu, and R. Zhang, "Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification," Electronics, vol. 11, no. 12, p. 1369, 2022.

[19] X. Li et al., "CANet: Cross-disease attention network for joint DR and DME grading," IEEE Trans. Med. Imaging, vol. 39, no. 5, pp. 1483–1493, 2020.

[20] M. R. Islam et al., "Applying supervised contrastive learning for the detection of diabetic retinopathy," Comput. Biol. Med., vol. 146, p. 105602, 2022.

[21] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[22] K. Van Leemput et al., "Automated model-based tissue classification of MR images of the brain," IEEE Trans. Med. Imaging, vol. 18, no. 10, pp. 897–908, 1999.

[23] A. M. Abdulghani et al., "Network-aware Gaussian mixture models for multi-objective SD-WAN controller placement," Electronics, vol. 14, no. 15, p. 3044, 2025.

[24] A. M. Abdulghani et al., "Dynamic multi-objective controller placement in SD-WAN: A GMM-MARL hybrid framework," Network, vol. 5, no. 4, p. 52, 2025.

[25] A. M. Abdulghani et al., "Enhancing Healthcare Network Effectiveness Through SD-WAN Innovations," in Studies in Systems, Decision and Control, vol. 233. Springer, Cham, 2025.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, 2018, pp. 3–19.

[27] J. D. Bodapati et al., "Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 10, pp. 9825–9839, 2023.

[28] M. Oulhadj et al., "Diabetic retinopathy prediction based on vision transformer and modified capsule network," Comput. Biol. Med., vol. 175, p. 108523, 2024.

[29] S. S. Mondal et al., "EDLDR: An ensemble deep learning technique for detection and classification of diabetic retinopathy," Diagnostics, vol. 13, no. 1, p. 124, 2023.

## BIOGRAPHIES OF AUTHORS

Aws H. Hamad received a B.Sc. in computer science from Mustansiriyah university.Then, he got a M.Sc. in Artificial intelligence from Hunan University, China in 2019. He has been working for the Iraqi Ministry of Higher Education and Scientific Research at Research and Development department. Recently, he has given lectures at university regarding Artificial lntelligence. He has some publications in international journals. His work interests include Artificial intelligence, Machine learning, Deep learning and Optimization. He can be contacted at email**:** aws.hamed@nahrainuniv.edu.iq