Comparison of HOG-based Classical Classifiers and CNN Models for Diabetic Retinopathy Diagnosis*

Yelizaveta Kabanova[1,*,†], Nataliia Kuznietsova[1,†], Kateryna Ivanko[1,†] and Vishwesh Kulkarni[2,3,†]

[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
[2] King's College London, London, UK
[3] SUSTech-King's School of Medicine, Shenzhen, China

Abstract

Diabetic Retinopathy (DR) is a major cause of vision loss in adults and requires accurate staging to ensure timely treatment. Existing studies on automated detection often neglect predictive confidence. The paper presents a comprehensive, confidence-aware comparison between classical classifiers (Support Vector Machines and Random Forests) based on Histogram of Oriented Gradients (HOG) and a state-of-the-art convolutional neural network (CNN) architecture with Focal Loss, Adaptive Convolutional Block Attention Modules (CBAM), and Monte Carlo Dropout for uncertainty estimation. Using the APTOS 2019 Blindness Detection dataset, we trained all models under identical preprocessing, data balancing, and evaluation protocols. Performance was evaluated using Quadratic Weighted Kappa (QWK), F1-score, AUC, and confidence-aware metrics such as AURC and E-AUROC. Results indicate that the CNN achieves the highest classification accuracy (QWK=0.92, F1=0.85, AUC=0.97) and most reliable uncertainty estimates, outperforming classical models. SVM combined with HOG remains competitive (QWK=0.85, F1=0.80), whereas Random Forest underperforms relative to both approaches. The findings highlight the importance of evaluating model reliability alongside conventional metrics in medical imaging and demonstrate that combining deep learning with uncertainty-aware strategies provides robust diagnostic support, while classical models retain value in resource-limited scenarios.

Keywords

diabetic retinopathy, fundus images, machine learning, image recognition, uncertainty estimation

CEUR-WS.org/Vol-4220/paper8.pdf

1. Introduction

Diabetic Retinopathy (DR) is a progressive retinal disease caused by chronic hyperglycemia that damages retinal blood vessels and remains one of the leading causes of vision loss among adults. Accurate staging is essential for timely treatment and prevention of irreversible vision loss [1].

Automated analysis of fundus images has become an important research direction. In recent years, deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated remarkable performance in retinal image classification.

Traditional machine learning methods, such as Support Vector Machines (SVMs) and Random Forest (RF), combined with handcrafted feature descriptors like Histogram of Oriented Gradients (HOG), remain computationally efficient and interpretable alternatives. However, researchers rarely evaluate them alongside modern CNN architectures under identical preprocessing procedures, class balancing strategies, and experimental conditions.

Another important limitation of existing studies is the insufficient analysis of predictive confidence. Most works rely on accuracy-based metrics, while in medical image analysis, the reliability of model predictions is equally critical. Incorrect but highly confident predictions may

* Computer Modeling and Intelligent Systems (CMIS-2026), May 05, 2026, Zaporizhzhia, Ukraine
[1,*] Corresponding author.
[†] These authors contributed equally.
yelizavetakabanova@gmail.com (Y. Kabanova); natalia-kpi@ukr.net (N. Kuznietsova); ivanko-ee@lll.kpi.ua (K. Ivanko); vishwesh.kulkarni@kcl.ac.uk (V. Kulkarni)
0009-0001-2692-5066 (Y. Kabanova); 0000-0002-1662-1974 (N. Kuznietsova); 0000-0002-3842-2423 (K. Ivanko); 0000-0002-22858652 (V. Kulkarni)
Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-07-04

lead to serious clinical consequences. Although researchers have applied uncertainty estimation techniques such as Monte Carlo Dropout to CNNs, classical machine learning classifiers can also provide confidence-related measures. Despite this, systematic comparison of predictive confidence across these fundamentally different paradigms remains largely unexplored.

This study performs a comprehensive, confidence-aware comparison between HOG-based classical classifiers (SVMs, RF) and a CNN architecture incorporating Focal Loss, Adaptive CBAM attention, and Monte Carlo Dropout. The main contributions of this study are as follows:

- A unified evaluation framework for a fair comparison under identical preprocessing, balancing, and training conditions.
- Extension of uncertainty analysis from Monte Carlo Dropout–based estimation to confidence–correctness alignment metrics (AURC, E-AURC).
- A reliability-centered interpretation of model behavior in diabetic retinopathy classification across fundamentally different learning paradigms.

2. Literature Review

Over the past few decades, numerous studies have investigated the automated diagnosis of diabetic retinopathy (DR) using classical machine learning techniques. Early research relied on handcrafted feature extraction combined with shallow classifiers.

Support Vector Machines (SVMs) were among the most widely adopted classifiers in early DR detection systems. Priya and Aruna [2] demonstrated that SVM-based classifiers outperform probabilistic neural networks (PNN) when applied to color fundus images. Similarly, Narasimhan et al. [3] reported that SVM outperforms Bayesian classifiers, achieving classification accuracy of up to 95%. Numerous subsequent studies implemented SVM-based methods for DR detection [4, 5, 6, 7]. However, most of these works focused on binary classification problems, without performing fine-grained grading of disease severity across multiple stages.

Random Forest (RF) classifiers were also extensively explored due to their ability to handle nonlinear decision boundaries and feature interactions. One of the earliest and most influential applications of RF methods to DR analysis was presented by Casanova et al. [8], who demonstrated the effectiveness of ensemble learning techniques for retinal image classification. Subsequent studies further investigated RF-based approaches and compared them with SVMs classifiers, reporting competitive performance across various datasets [9, 10, 11].

Despite promising performance, classical ML methods strongly depend on the quality of handcrafted features and preprocessing pipelines. Moreover, their generalization across datasets collected under different imaging conditions is limited, restricting clinical scalability.

To improve classical ML performance, extensive research has focused on handcrafted feature extraction methods, particularly those based on localized retinal lesions such as microaneurysms, hemorrhages, and exudates. Histogram of Oriented Gradients has been widely adopted for capturing edge and structural information in fundus images. Sarwinda et al. [12] proposed a DR classification framework based on HOG features combined with shallow learning classifiers, including SVM and Random Forest, reaching approximately 85% in multi-class DR classification tasks. Gandor et al. [13] combined Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrices (GLCM) features with Random Forest classifiers, achieving an accuracy of 80.41% and an AUC of 0.80 on fundus image datasets.

The introduction of deep learning, particularly convolutional neural networks, eliminated the need for manual feature engineering and enabled end-to-end learning directly from raw images. Gulshan et al. [14] demonstrated that deep CNN models trained on large-scale retinal datasets can achieve diagnostic performance comparable to, and in some cases exceeding, that of expert ophthalmologists. Subsequent studies adopted pretrained architectures and explored various strategies to enhance DR classification, including attention mechanisms, specialized loss functions, and advanced image preprocessing techniques [15, 16, 17].

Recent studies have explored hybrid approaches that combine handcrafted features with deep learning representations. Ahmed [18] proposed a HOG-CNN framework that integrates texture-based HOG features with CNN embeddings for retinal image classification. The proposed method achieved high accuracy and AUC on multiple benchmark datasets, including APTOS 2019.

However, many existing studies neglect critical aspects required for real-world clinical deployment. Class imbalance is often insufficiently addressed, advanced preprocessing pipelines are not consistently applied, and predictive uncertainty is rarely evaluated. Existing uncertainty-aware research predominantly focuses on deep learning models and rarely compares uncertainty behavior across fundamentally different machine learning approaches [19].

This work addresses these gaps by conducting a comprehensive comparison between HOG-based classical classifiers (SVM, RF) and a deep learning architecture incorporating advanced preprocessing, oversampling-based class balancing, adaptive attention mechanisms, and Monte Carlo Dropout–based uncertainty estimation. In addition to performance evaluation, we analyze predictive confidence across models. All experiments were performed on the APTOS 2019 Blindness Detection dataset using consistent evaluation metrics to ensure a fair comparison of performance and reliability.

3. Materials and Methods

This study follows an experimental pipeline designed to ensure a fair comparison between classical machine learning and deep learning approaches for diabetic retinopathy classification. The workflow consists of four main stages: image preprocessing, class balancing and data augmentation, model training, and confidence-aware evaluation.

3.1. Image Preprocessing

Robust preprocessing is crucial for medical image analysis, as most fundus images exhibit uneven lighting, black borders, artifacts, glare, and color variations due to different imaging conditions and equipment. Some images contain the entire retina, but most of them do not have the upper and lower segments. These regions do not contain useful information, but can affect the pixel intensity distribution and stability of the model training. Our preprocessing pipeline included the following steps:

1. Cropping: black corners from images were removed using an algorithm based on pixel-intensity thresholding [21].
2. Resizing: cropped images were resized to 224×224 pixels.
3. Normalization: pixel intensities were scaled to the [0,1] range.
4. Gaussian blurring: a Gaussian filter was applied to reduce high-frequency noise while preserving relevant structural information.
5. Linear enhancement: blurred and original images were linearly combined to enhance vessel structures and contrast
6. Contrast enhancement: CLAHE (Contrast Limited Adaptive Histogram Equalization) was applied to improve visibility of subtle pathological patterns.

Figure 1 illustrates the preprocessing steps.

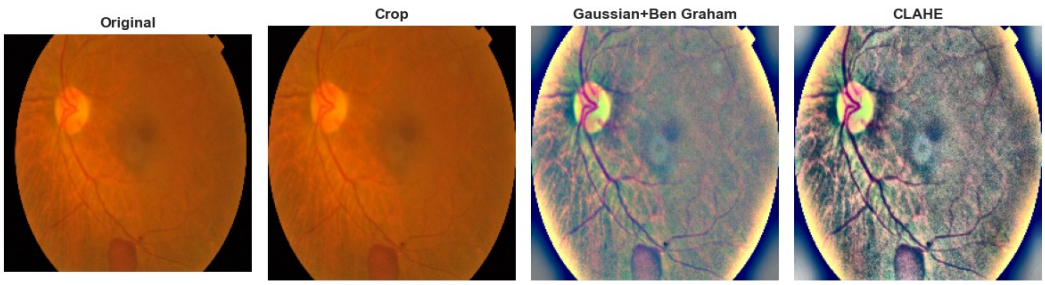


Figure 1: Visualization of image preprocessing steps.

This preprocessing improves the visibility of small retinal details, which is critical for accurate DR classification.

3.2. Class Balancing Strategy

In the APTOS 2019 BD dataset, class distribution is highly imbalanced, with class 0 dominating 49.3% of samples (1805 images), followed by class 2 with 27.3% (999), while classes 1, 3, and 4 are underrepresented with 10.1% (370), 5.3% (193), and 8.1% (295), respectively.

To mitigate this imbalance, we applied oversampling for minority classes using custom augmentations, increasing the frequency of underrepresented classes to match the size of the majority class (N=1805). After balancing, the training set contained 9,025 images. This approach improved model generalization while preserving the original class distribution outside the balancing process.

3.3. Data Augmentation

To enhance model robustness and generalization to unseen images, we applied two levels of data augmentation:

- Balancing stage: during oversampling, additional images were generated for minority classes using horizontal flips, brightness adjustments (±20%), and rotations (±15°).
- Training stage: additional transformations, including zoom (±10%), width and height shifts (±10%), along with horizontal flips, brightness, and rotation (±10°), were applied to the training set.

3.4. Deep Learning Model Architecture

In this study, the configuration of the best-performing model developed in the preliminary experiments was selected for comparison with classical machine learning approaches.

The proposed architecture is based on transfer learning with a DenseNet121 backbone pretrained on ImageNet. During training, the convolutional base was initially frozen and subsequently partially fine-tuned by unfreezing the last 20 layers. Additionally, batch normalization was applied to stabilize feature distributions and improve convergence during training.

To enhance the representational capacity of the network, Adaptive Convolutional Block Attention Modules (Adaptive CBAM) were integrated after the convolutional backbone. The module structure consists of two sequential sub-modules: channel attention and spatial attention. These steps help to identify which features are important and then locate them on the retinal image. The Adaptive Module dynamically adjusts both channel and spatial attention based on feature variance, enabling the model to focus on clinically relevant retinal structures such as microaneurysms, hemorrhages, and exudates. Global average pooling was then applied to aggregate spatial information and reduce the number of trainable parameters.

The classification head consists of a fully connected layer with 256 neurons and ReLU activation with L2 regularization. To address epistemic uncertainty and examine prediction reliability, Monte Carlo Dropout (rate=0.3) was incorporated during inference by performing multiple stochastic passes for each input image to obtain a predictive distribution.

Given the severe class imbalance, the network was optimized using Focal Loss, which reduces the contribution of easily classified examples and enhances the influence of complex examples during training. This helps increase the sensitivity of the model to severe stages of the disease.

Training used Adam with learning rates 0.0001 (head) and 0.00001 (fine-tuning), batch size 32, 30 epochs, fixed seed.

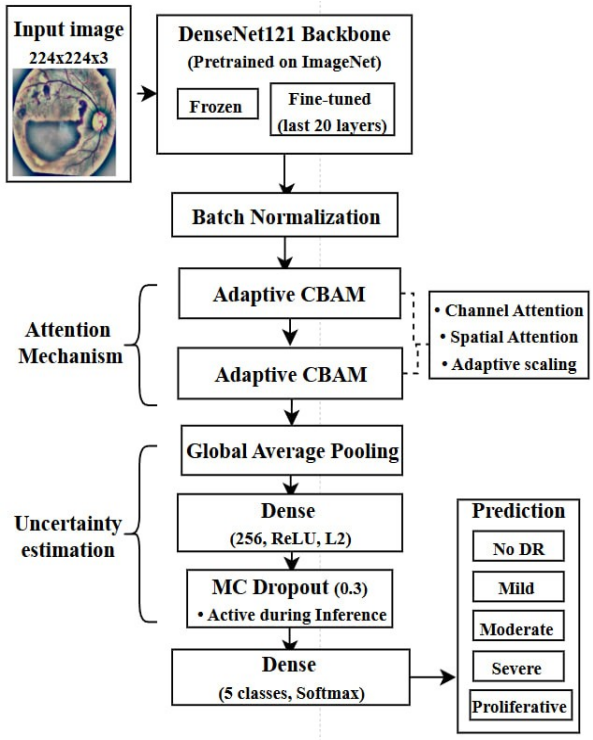


Figure 2: Proposed deep learning architecture for diabetic retinopathy classification.

This study introduces a unified confidence-aware evaluation framework that enables a systematic comparison between deep learning and classical machine learning approaches under identical preprocessing, balancing, and training conditions.

3.5. Handcrafted Feature Extraction Using HOG

Histogram of Oriented Gradients (HOG), introduced by Navneet Dalal and Bill Triggs [22], encodes local structural information by analyzing the distribution of gradient orientations within small spatial regions. Due to its ability to capture edges, contours, and texture patterns, HOG is well-suited for representing retinal structures in fundus images.

Each preprocessed fundus image was converted to a grayscale image to allow the descriptor to focus on intensity-based structural information, such as vessels, lesions, and texture variations in retinal images. Image gradients in both horizontal and vertical directions were then calculated using discrete derivative masks:

$$m(x,y)=\sqrt{G_x^2+G_y^2}, \tag{1}$$

$$\theta(x,y)=\arctan\left(\frac{G_y}{G_x}\right), \tag{2}$$

where (x, y) is a spatial location, $m(x, y)$ denotes gradient magnitude, $\theta(x, y)$ is a gradient orientation, and G_x and G_y denote the horizontal and vertical image gradients, respectively.

The image was then partitioned into small spatial regions called cells. Within each cell, a histogram of gradient orientations is constructed by accumulating gradient magnitudes into a fixed number of orientation bins. These histograms represent the local distribution of edge directions and encode texture and shape information. Several parameter configurations were evaluated to determine the most suitable representation for retinal structures:

- Number of orientation bins: 9, 12.
- Pixels per cell: 4×4, 6×6, 8×8.
- Cells per block: 2×2, 3×3.

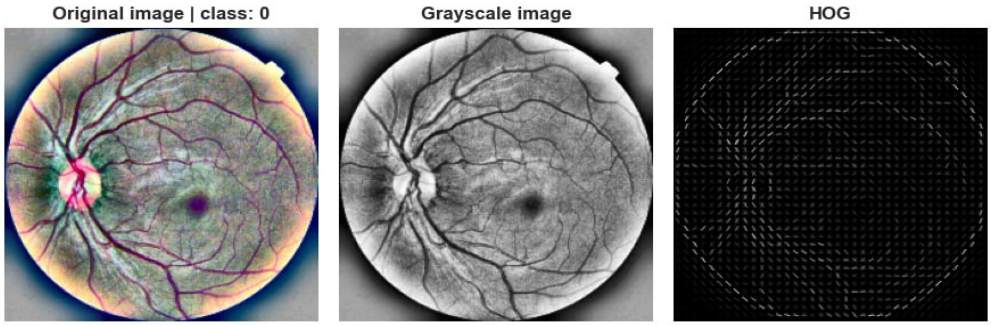
To improve robustness and reduce the influence of extreme intensity values, histograms were normalized using the L2-Hys normalization method, which limits the maximum value of the descriptor vector for each block to a predefined threshold [22].

The final feature vector was formed as:

$$F_k = HOG(G_k; o=9, p=6\times6, b=2\times2), \quad (3)$$

where G_k denotes the grayscale image, o is the number of orientation bins, p is the cell size, and b is the block size.

These feature vectors were subsequently used as inputs for classical machine learning classifiers. Figure 3 illustrates the transformation pipeline.



a) b) c)

Figure 3: Illustration of the feature extraction pipeline applied to a fundus image: (a) preprocessed RGB fundus image, (b) grayscale conversion used for HOG computation, and (c) corresponding HOG representation highlighting edge and texture features.

Unlike convolution neural networks, which automatically learn hierarchical features directly from data, HOG relies on manually designed rules for feature extraction based on image gradients. HOG-based approaches are more interpretable and computationally efficient than CNN architectures, making them suitable for resource-constrained scenarios. However, the performance of HOG-based methods strongly depends on careful parameter selection and preprocessing quality [23].

HOG provides a strong handcrafted feature representation that significantly enhances the performance of classical machine learning models and serves as a meaningful baseline for comparison with modern deep learning approaches.

3.6. Classical Machine Learning Classifiers

Support Vector Machine (SVM) is a classical supervised machine learning algorithm developed by Cortes and Vapnik [24]. The primary objective of SVM is to construct a decision boundary that maximizes the margin between classes, improving generalization on unseen data. Unlike earlier classifiers that focused solely on minimizing training error, SVM explicitly optimizes the trade-off between margin maximization and classification error.

Given the set of training samples (x_i, y_i), where $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$, a linear SVM aims to find an optimal separating hyperplane defined as:

$$w^T x + b = 0, \tag{4}$$

by solving the optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i, \tag{5}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \tag{6}$$

where w is the normal vector to the hyperplane, b is the bias, ξ_i are slack variables allowing misclassification, and C is a regularization parameter controlling the trade-off between margin width and classification error.

To handle non-linearly separable data, SVM employs kernel functions that implicitly map input data into a higher-dimensional feature space, where linear separation becomes feasible [24, 25]. In this study, radial basis function (RBF) was selected due to its effectiveness for complex, high-dimensional feature spaces, common for medical images. For multi-class diabetic retinopathy grading, the One-vs-One (OvO) strategy was applied, constructing binary classifiers for each pair of classes to have more balanced decision boundaries and reduce bias toward dominant classes.

Support Vector Machine has been applied in diabetic retinopathy classification due to its robustness in high-dimensional spaces and its ability to perform well on limited or imbalanced datasets. Combined with HOG, SVM has demonstrated competitive performance in DR detection and grading tasks, remaining a suitable baseline for comparison with deep learning approaches.

Random Forest (RF) is an ensemble machine learning method introduced by Breiman [26], which constructs a collection of decision trees by randomly selecting samples and feature subsets. The core idea of Random Forest is to reduce variance and improve generalization by aggregating multiple weak learners into a strong predictor.

Formally, an RF consists of tree-structured classifiers, each trained on a bootstrap sample drawn independently from the original training set. At each node, a random subset of features is considered for splitting, increasing model diversity [26, 27]. For multi-class classification tasks, the RF prediction is obtained via majority voting:

$$\hat{y}(x) = \underset{c \in C}{\operatorname{argmax}} \sum_{m=1}^{M} I(h_m(x) = c), \tag{7}$$

where $h_m(x)$ denotes the class predicted by the tree m for input x, C is the set of class labels, and I is the indicator function.

Unlike SVM, Random Forest is less sensitive to noise in individual features, making it a strong ML method for comparison. In this study, RF was used with HOG feature vectors, and the influence of the number of trees on model stability and performance was evaluated for sizes of 200, 300, and 500.

3.7. Experimental Setup

Since HOG produces high-dimensional feature vectors with different value ranges across components, the Standard Scaler was applied before classification.

For dimensionality reduction and noise suppression, Principal Component Analysis (PCA) was applied after standardization. To determine the optimal number of principal components for the HOG features, multiple experiments were conducted. The number was selected based on the highest Quadratic Weighted Kappa (QWK) achieved on the validation subset, resulting in 160 components.

To investigate the influence of dimensionality reduction on ensemble learning, Random Forest configurations included an experiment with and without PCA. This comparison allows analysis of how feature space compression affects tree-based models, which are less sensitive to feature scaling than SVMs.

3.8. Performance Evaluation Metrics

In this study, the analysis of the results is performed using classical machine learning metrics, metrics specifically designed for imbalanced data, and predictive confidence evaluation. The metrics included:

- Accuracy:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

- Precision and Recall:

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}. \tag{10}$$

- F1-score:

$$F1 = \frac{2 \cdot Precison \cdot Recall}{Precison + Recall}, \tag{11}$$

where TP, TN, FP, FN denotes true positive, true negative, false positive, and false negative values, respectively.

- Quadratic Weighted Kappa (QWK):

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \tag{12}$$

where O_{ij} is the observed confusion matrix, E_{ij} is the expected matrix under random agreement, and w_{ij} is the quadratic weight defined as:

$$w_{ij} = \frac{(i-j)^2}{(C-1)^2}, \tag{13}$$

- Area Under the ROC Curve (AUC).

Among these metrics, QWK was chosen as the main comparison metric, as it takes into account the ordinal nature of the classes and the different error weights between neighboring stages of the disease.

Traditional evaluation metrics measure the accuracy of the model, indicating how often a model is correct, but they do not analyze whether the model knows when it might be wrong. Model reliability was evaluated using confidence-aware measures:

- For the CNN model, confidence corresponded to the mean predictive probability obtained via Monte Carlo Dropout sampling.
- For the SVM model, prediction confidence was computed from an auxiliary One-vs-Rest (OvR) SVM, as the maximum predicted class probability.
- For RF models, confidence was defined as the maximum predicted class probability.

From these confidence values, the following metrics were computed [28]:

- AURC (Area Under Risk–Coverage Curve) evaluates the quality of uncertainty estimates by measuring how prediction risk changes when low-confidence samples are gradually rejected. Lower AURC indicates that most errors are concentrated among low-confidence predictions, meaning the model can maintain high accuracy on confident cases while deferring uncertain samples.
- Error AURC (E-AURC) measures how well confidence distinguishes correct from incorrect predictions. Higher values indicate better ability to detect potential errors.
- Confident Separation quantifies a difference between the mean confidence of correct and incorrect predictions.

Together, these metrics provide insight into how well model confidence correlates with prediction correctness. Confidence scores are treated as relative ranking values rather than strictly comparable calibrated probabilities across different model families.

3.9. Dataset

In this study, we used the APTOS Blindness Detection 2019 (APTOS 2019 BD) [20], which was created for a Kaggle competition on diabetic retinopathy (DR) diagnosis. The dataset contains 3,662 color fundus images and was provided by Aravind Eye Hospital, India, representing the Asia Pacific Tele-Ophthalmology Society. Each image was manually labelled into five classes (0–4), indicating DR severity.

To ensure reliable performance evaluation and proper generalization of the proposed models, the original dataset was divided into training and validation subsets using an 80/20 split ratio. The training subset contained 7,220 images after class balancing, while the validation subset contained 1,805 images.

4. Experimental Results

The quantitative results for all evaluated configurations are summarized in Table 1.

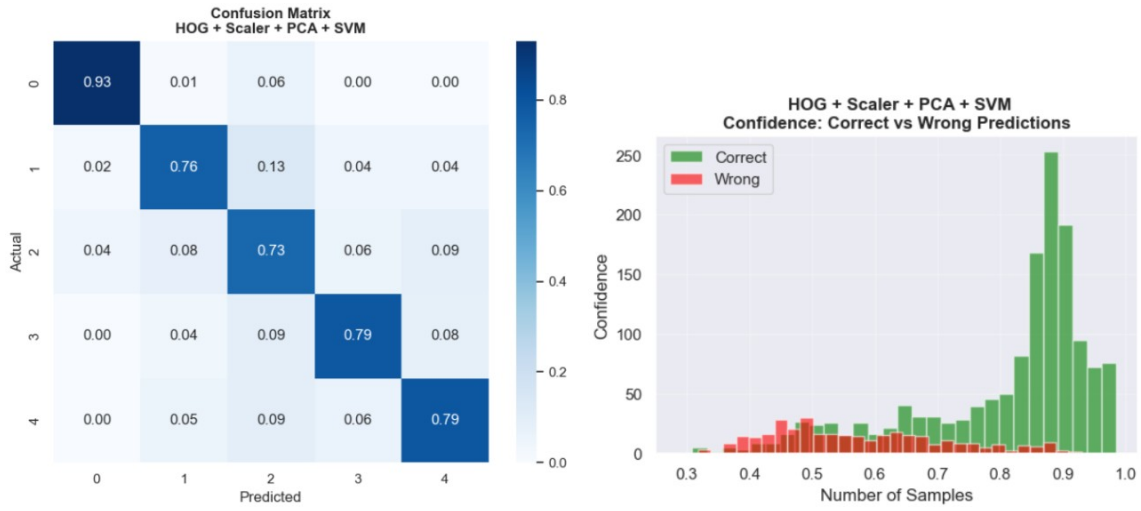
Table 1
Classification results for all evaluated configurations

Model		QWK	ACC	Prec.	Rec.	F1	AUC	AURC	E-AURC	Mean Conf.	Conf. Separation
HOG + Scaler + PCA + SVM		**0.851**	**0.799**	**0.802**	**0.799**	**0.8**	**0.930**	**0.054**	**0.862**	**0.76**	**0.23**
HOG + RF	200	0.735	0.708	0.717	0.708	0.707	0.92	0.11	0.825	0.52	0.251
	300	**0.739**	**0.713**	**0.723**	**0.713**	**0.712**	**0.924**	**0.108**	**0.817**	**0.519**	**0.248**
	500	0.737	0.709	0.720	0.709	0.708	0.926	0.106	0.829	0.517	0.254
HOG + PCA + RF	200	0.784	0.760	0.764	0.760	0.759	0.935	0.085	0.828	0.546	0.246
	300	**0.784**	**0.761**	**0.766**	**0.761**	**0.761**	**0.937**	**0.084**	**0.828**	**0.545**	**0.247**
	500	0.784	0.761	0.766	0.761	0.761	0.937	0.084	0.828	0.545	0.247
CNN DenseNet 121		**0.922**	**0.856**	**0.917**	**0.735**	**0.855**	**0.979**	**0.033**	**0,84**	**0,75**	**0.227**

The developed **CNN** configuration achieved the highest overall classification performance among all evaluated models. It achieved a Quadratic Weighted Kappa (QWK) score of 0.92, an accuracy of 0.85, and the highest AUC value of 0.97. Additionally, the model demonstrated the lowest AURC (0.03), indicating strong reliability of predictions under uncertainty-based ranking. The model also maintained a balanced confidence and achieved strong class-wise performance, including minority classes. These findings confirm the robustness of the deep learning approach for the given classification task.

The SVM model demonstrated the strongest performance among traditional machine learning approaches. The model achieved a QWK of 0.85, an accuracy of 0.80, and an F1-score of 0.80. According to the confusion matrix (Figure 4), the dominant class (class 0) achieved the highest classification accuracy (0.94), while the remaining classes achieved scores below 0.80, indicating class imbalance sensitivity. In contrast, the CNN achieved consistently higher classification accuracy across all classes.

In terms of uncertainty estimation, the SVM model showed a mean confidence of 0.76, marginally higher than that of the CNN. Moreover, it achieved the highest E-AUROC (0.86) among all models, indicating strong separation between correct and incorrect predictions based on confidence scores. The distribution of confidence for correct and incorrect predictions is shown in Figure 4.



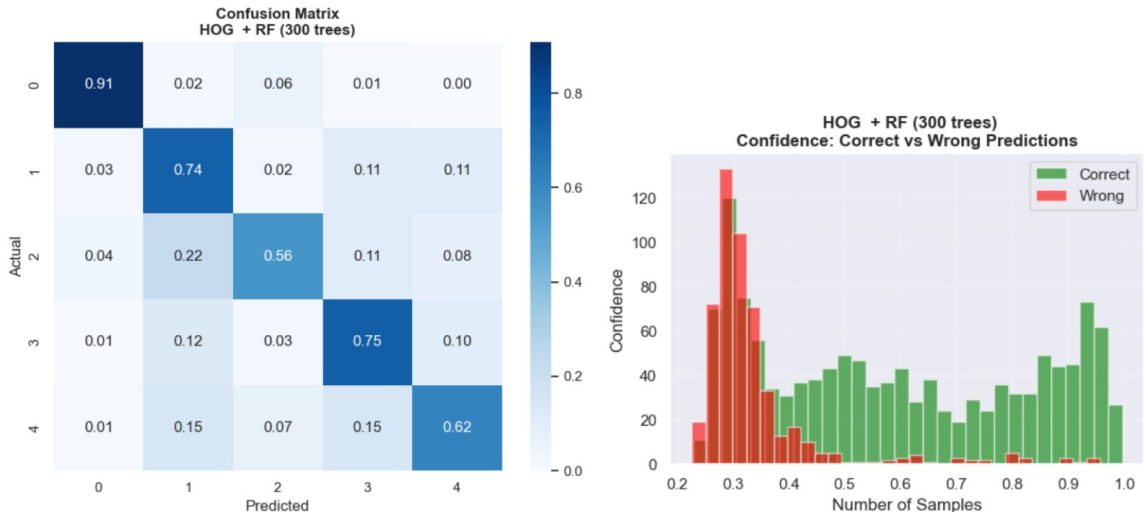
a) b)

Figure 4: (a) Confusion matrix of the SVM model and (b) confidence distribution for correct and incorrect predictions.

Random Forest was first evaluated without dimensionality reduction. The results were similar across different numbers of trees (200, 300, and 500), with slightly better performance for the configuration with 300 trees. Performance metrics remained in the range of:

- QWK: ~0.73–0.74.
- Accuracy: ~0.71–0.72.
- F1-score: ~0.71.
- AUC: ~0.92–0.93.

The confusion matrix for the best-performing configuration without PCA and the corresponding confidence distribution is shown in Figure 5. Mean confidence values were approximately 0.51, which is notably lower than those of both SVM and CNN. The E-AUROC values (~0.82–0.83) indicate weaker separation between correct and incorrect predictions than those of SVM. Furthermore, the AURC values (~0.11) are considerably higher than those of CNN and SVM.



a) b)

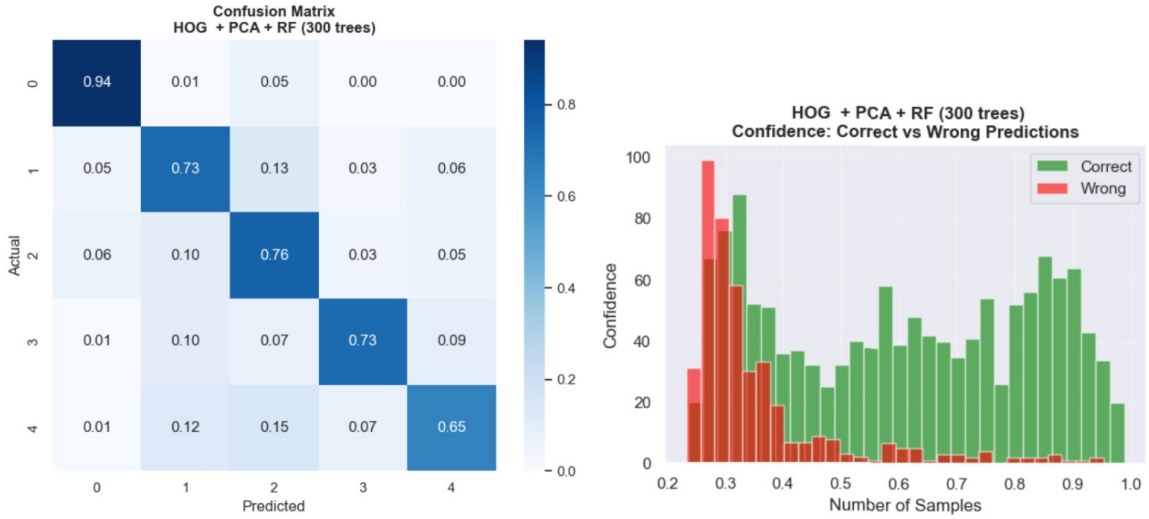
Figure 5: (a) Confusion matrix of the RF model without PCA and (b) confidence distribution for correct and incorrect predictions.

Applying PCA before Random Forest notably improved the performance of classification, although variations in the number of trees again did not lead to significant changes. The performance increased to approximately:

- QWK: ~0.78.
- Accuracy: ~0.76-0.77.
- F1-score: ~0.76.
- AUC: ~0.94.

The confusion matrix for the best-performing PCA-based RF configuration and the corresponding confidence distribution is presented in Figure 6. The most noticeable improvements were observed in uncertainty estimation. The mean confidence increased slightly to approximately 0.55, compared to RF without PCA. The E-AUROC values (~0.83) slightly improved, while the AURC values (~0.08) decreased compared to the non-PCA RF model, indicating more reliable uncertainty ranking.

However, despite these improvements, Random Forest with PCA still underperformed compared to both SVM and CNN across nearly all evaluation metrics.



a) b)

Figure 6: (a) Confusion matrix of the RF model with PCA and (b) confidence distribution for correct and incorrect predictions.

5. Discussion of results

The results of this study demonstrate that the CNN model outperformed classical approaches in both classification performance and uncertainty-aware evaluation. This superiority suggests that learned hierarchical feature representations provide a substantial advantage over handcrafted descriptors such as HOG. Although HOG captures local gradient structures, it remains limited in representing complex spatial patterns and class-specific variations, and is less effective in handling class imbalance.

In contrast, the proposed CNN with an integrated adaptive module, focal loss function, and Monte Carlo Dropout achieved significant improvements in QWK, AUC, and class-wise stability. Furthermore, the CNN model demonstrated not only the highest classification performance, but also the best uncertainty-related metrics, particularly the lowest AURC. This indicated that deep learning architectures are better suited to decision-support scenarios in medical applications, where unreliable predictions must be identified and potentially rejected.

The SVM model demonstrated that classical machine learning approaches can remain competitive when combined with strong handcrafted feature engineering. Its high E-AUROC suggests that SVM confidence values separate correct and incorrect predictions effectively. However, its low class-wise performance indicates that margin-based classifiers struggle when under class imbalance.

The Random Forest results remained weaker than those of SVM and CNN, despite an increasing number of trees. It was also discovered that the introduction of PCA led to moderate improvements, suggesting that dimensionality reduction helped remove noise.

Another key finding is the relationship between confidence magnitude and confidence reliability. RF models produced lower mean confidence but also weaker separation between correct and incorrect predictions. In contrast, SVM showed higher confidence and stronger separability, while CNN achieved a better balance between confidence level and reliability. This suggests that not only confidence magnitude but also its distribution relative to correctness determines the effectiveness of uncertainty-based decision strategies.

Overall, the findings support the view that not only accuracy of predictions, but confidence in predictions should be considered when selecting a baseline for classification framework. The study demonstrates that deep learning models offer superior feature learning and more informative confidence estimation than classical machine learning models. Nevertheless, the obtained results also indicate that the HOG + SVM configuration remains a robust option.

6. Conclusions

This study introduced a unified confidence-aware framework for the comparative evaluation of deep learning and classical machine learning models in diabetic retinopathy classification. By extending uncertainty analysis beyond Monte Carlo Dropout to confidence–correctness alignment metrics, the work provides a reliability-centered perspective on model performance under a controlled experimental setup. Within the work, we presented a comprehensive comparison between handcrafted classical machine learning classifiers using HOG and a deep learning architecture for multi-class diabetic retinopathy classification under identical preprocessing, balancing, and evaluation conditions, with particular attention to predictive confidence and reliability.

The results demonstrate a clear advantage of deep learning architecture for both classification performance and uncertainty-aware evaluation. The proposed CNN architecture achieved the highest overall performance (QWK = 0.92, F1 = 0.85, AUC = 0.97) together with the lowest AURC, indicating superior reliability of predictions under confidence-based ranking.

Among classical approaches, the HOG combined with the SVM configuration demonstrated the strongest performance (QWK = 0.85, F1 = 0.80, AUC = 0.93), confirming that well-designed handcrafted features remain competitive. However, class-wise analysis revealed reduced performance for minority classes, indicating that margin-based classifiers remain sensitive to class imbalance despite preprocessing and balancing strategies.

Random Forest models demonstrated lower classification performance and less informative confidence estimates compared to both CNN and SVM. Dimensionality reduction using PCA improved stability and reliability metrics, while increasing the number of trees did not lead to significant changes. The best configuration achieved QWK = 0.78, F1 = 0.76, AUC = 0.93.

Another important finding of this study is the critical role of confidence–correctness alignment in reliable medical prediction. It was shown that uncertainty estimation is not determined solely by confidence, but by how well confidence correlates with prediction correctness. The CNN model demonstrated the best balance between confidence level and reliability, while SVM showed strong error–confidence separability, and Random Forest produced weaker uncertainty discrimination.

A limitation of this study is that the proposed framework was evaluated solely on the APTOS 2019 dataset. External validation on other benchmark datasets, such as EyePACS or

Messidor, was not performed. Therefore, future research should focus on validating the framework on larger and more diverse datasets to assess its generalization capabilities across different clinical conditions and imaging equipment. Furthermore, investigation of alternative uncertainty estimation methods, advanced class balancing techniques, and various adaptive attention mechanisms may further improve model robustness and reliability in real-world clinical deployment.

Overall, the study confirms that confidence-aware evaluation provides essential information beyond traditional accuracy-based metrics in medical image analysis. Deep learning models offer superior diagnostic performance and more reliable confidence estimation, making them well-suited for clinical decision-support systems. Nevertheless, the HOG combined with SVM approach remains a viable and computationally efficient baseline for resource-limited environments.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.2 and Grammarly in order to: Text Translation, Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed, and took full responsibility for the publication's content.

References

[1] F. Bandello, M. A. Zarbin, R. Lattanzio, I. Zucchiatti, Clinical Strategies in the Management of Diabetic Retinopathy: A Step-By-step Guide for Ophthalmologists, Springer, London, 2014.

[2] R. P. R. Priya, P. Aruna, SVM and neural network based diagnosis of Diabetic Retinopathy, Int. J. Comput. Appl. 41 (2012) 6–12. doi:10.5120/5503-7503.

[3] K. Narasimhan, V. C. Neha, K. Vijayarekha, An efficient automated system for detection of diabetic retinopathy from fundus images using support vector machine and bayesian classifiers, in: Proceedings of the International Conference on Computing, Electronics and Electrical Technologies, IEEE, 2012, pp. 964–969. doi:10.1109/ICCEET.2012.6203804.

[4] R. Sahebrao, S. N., S. T., M. Dhopeshwarkar, Automated Diagnosis Non-proliferative Diabetic Retinopathy in Fundus Images using Support Vector Machine, Int. J. Comput. Appl. 125 (2015) 7–10. doi:10.5120/ijca2015905968.

[5] M. Gandhi, D. Raghavan, Diagnosis of diabetic retinopathy using morphological process and SVM classifier, in: Proceedings of the International Conference on Communication and Signal Processing, IEEE, 2013, pp. 873-877. doi:10.1109/iccsp.2013.6577181.

[6] E. Shahin, T. Taha, W. Al-Nuaimy, E. El-Rabaie, O. Zahran, F. Abd El-Samie, Automated detection of diabetic retinopathy in blurred digital fundus images, in: Proceedings of the International Conference on Engineering Nano Communication and Optimization, IEEE, 2012, pp. 20-25. doi:10.1109/ICENCO.2012.6487084.

[7] A. Vyas, et al., Detection of diabetic retinopathy in fundus images using SVM, Int. J. Sci. Res. Eng. Manag. 6 (2022).

[8] R. Casanova, et al., Application of random forest methods to diabetic retinopathy classification analyses, PLoS ONE 9 (2014) e98587. doi:10.1371/journal.pone.0098587.

[9] N. Abhiram, M. B. Joseph, A. Sharmila, Detection of diabetic retinopathy using SVM and random forest classifiers, in: Proceedings of the International Conference on Intelligent Paradigms and Applications in Technology (I-PACT), IEEE, 2023. doi:10.1109/I-PACT58649.2023.10434824.

[10] A. M. Nasir, et al., Automatic diabetic retinopathy detection using random forest classifier, in: Proceedings of the IEEE TENCON, IEEE, 2022. doi:10.1109/TENSYMP54529.2022.9864355.

[11] I. Bhattacharjee, et al., Diabetic retinopathy classification using machine learning approaches, arXiv:2412.02265 (2024). doi:10.48550/arXiv.2412.02265.

[12] D. Sarwinda, T. Siswantining, A. Bustamam, Classification of diabetic retinopathy stages using HOG and shallow learning, in: Proceedings of the International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2018. doi:10.1109/IC3INA.2018.8629502.
[13] M. Gandor, et al., Diagnostics of diabetic retinopathy using machine learning with advanced feature engineering, Sci. Rep. 15 (2025) 34486. doi:10.1038/s41598-025-06973-z.
[14] V. Gulshan, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (2016) 2402–2410. doi:10.1001/jama.2016.17216.
[15] M. R. Basarab, K. O. Ivanko, Investigation of fundus images for diabetic retinopathy stage detection using deep learning, Visnyk NTUU KPI (2023) 49–57. doi:10.20535/radap.2023.94.49-57.
[16] M. S. Harisha, A. A. Bhosale, Detection and classification of diabetic retinopathy using deep learning algorithms, arXiv:2401.02759 (2024). doi:10.48550/arXiv.2401.02759.
[17] N. Tsiknakis, D. Theodoropoulos, G. Manikis, et al., Deep learning for diabetic retinopathy detection and classification based on fundus images: A review, Comput. Biol. Med. 135 (2021) 104599. doi:10.1016/j.compbiomed.2021.104599.
[18] F. Ahmed, HOG-CNN: Integrating handcrafted and deep features for retinal classification, arXiv:2507.22274 (2025). doi:10.48550/arXiv.2507.22274.
[19] M. Akram, et al., Uncertainty-aware diabetic retinopathy detection using Bayesian deep learning, Sci. Rep. 15 (2025) 1342. doi:10.1038/s41598-024-84478-x.
[20] Kaggle, APTOS 2019 Blindness Detection Dataset, 2019. URL: https://kaggle.com/c/aptos2019-blindness-detection.
[21] N. Sikder, M. S. Chowdhury, A. S. M. Arif, A.-A. Nahid, Early blindness detection based on retinal images using ensemble learning, arXiv:2006.07475 (2020). URL: https://arxiv.org/pdf/2006.07475.
[22] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Diego, 2005, pp. 886–893. doi:10.1109/CVPR.2005.177.
[23] A. Suleiman, Y.-H. Chen, J. Emer, V. Sze, Towards closing the energy gap between HOG and CNN features for embedded vision, arXiv:1703.05853 (2017). doi:10.48550/arXiv.1703.05853.
[24] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[25] T. Evgeniou, M. Pontil, Support vector machines: Theory and applications, in: Machine Learning and Its Applications, Springer, Berlin, 2001, pp. 249–257. doi:10.1007/3-540-44673-7_12.
[26] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[27] G. Biau, E. Scornet, A random forest guided tour, TEST 25 (2016) 197–227. doi:10.1007/s11749-016-0481-7.
[28] Y. Geifman, G. Uziel, R. El-Yaniv, Bias-reduced uncertainty estimation for deep neural classifiers, in: Proceedings of the International Conference on Learning Representations (ICLR), 2019.