IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3745~3760
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3745-3760

Adaptive binary capsule network for complex image recognition

Mavis Serwaa Yeboah[1,2], Patrick Kwabena Mensah[1], Adebayo Felix Adekoya[3], Mighty Abra Ayidzoe[1,4]
[1]Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana
[2]Department of Computer Science, Sunyani Technical University, Sunyani, Ghana
[3]Faculty of Computing, Engineering and Mathematical Sciences, Catholic University of Ghana, Sunyani, Ghana
[4]School of Technology, Jamk University of Applied Sciences, Jyväskylä, Finland

Article Info

Article history:

Received Dec 5, 2024
Revised Jun 3, 2026
Accepted Jul 9, 2026

Keywords:

Adaptive contrast
Capsule network
Complex images
Convolutional neural networks
Local binary pattern

ABSTRACT

Glaucoma and cataracts are leading causes of blindness worldwide, emphasizing the need for early detection. Convolutional neural networks (CNNs) have shown promise in detecting ocular diseases, but require extensive training datasets. However, medical datasets are scarce, limited, and imbalanced, prompting the use of time-consuming data augmentation techniques. To address these limitations, a computationally efficient and robust capsule network (CapsNet) model was proposed. The model features a novel adaptive contrast spatial filtering (ACSF) algorithm and incorporates a local binary pattern (LBP) algorithm to enhance robustness. This study achieved good recognition accuracy, with scores of 96.90% on the combined dataset, 96.45% on the cataract-only dataset, and 96.78% on the glaucoma-only dataset. The model's performance is comparable to state-of-the-art models, demonstrating its potential to support ophthalmologists in diagnosing cataracts and glaucoma-related eye issues.

Corresponding Author:

Mavis Serwaa Yeboah
Department of Computer Science and Informatics, University of Energy and Natural Resources
Sunyani, Ghana
Email: mavis.serwaayeboah@stu.edu.gh

1. INTRODUCTION

Glaucoma represents a critical global health challenge as a leading cause of irreversible blindness, characterized by progressive optic nerve degeneration that, unlike cataracts, cannot be reversed through surgical intervention. With projections estimating that 111 million people will be affected by glaucoma by 2040 [1], early detection has become paramount to preventing permanent vision loss. This urgency has driven the development of intelligent screening and diagnostic systems capable of identifying ocular abnormalities before irreversible damage occurs. However, the effectiveness of these systems depends fundamentally on their ability to accurately distinguish between multiple ocular conditions while operating within the constraints of real-world clinical settings.

Deep learning algorithms, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in identifying ocular abnormalities from retinal images. Recent advances include hybrid CNN-adaptive boosting (AdaBoost) models achieving 92.96% accuracy on glaucoma datasets [2], transfer learning approaches with pre-trained architectures (residual network 50 (ResNet-50), visual geometry group 19 (VGG-19), densely connected convolutional network 201 (DenseNet-201)) yielding accuracies up to 99.57% [3], and fine-tuned ResNet-50 models reaching 98.48% accuracy through data augmentation strategies [4]. Despite these achievements, CNNs face significant challenges related to data scarcity and quality. Medical image datasets are typically limited in size and characterized by class imbalance, leading to overfitting and reduced generalization capability [5]. While data augmentation techniques partially address

these limitations [5], they do not fundamentally resolve the architectural constraints of CNNs in capturing spatial hierarchies and part-whole relationships in medical images.

Capsule networks (CapsNets) have emerged as a promising alternative architecture that retains CNN strengths while addressing data-related challenges through their unique ability to preserve spatial relationships and adapt to smaller datasets. Initial applications to ocular disease diagnosis have shown encouraging results: the original CapsNet with dynamic routing achieved 90.90% accuracy on glaucoma datasets [6], modified 3D CapsNets attained 0.89 sensitivity on optical coherence tomograph (OCT) images [7], and enhanced CapsNets with feature processing algorithms demonstrated precision values of 0.796 for glaucoma and 0.818 for cataracts. However, CapsNets remain relatively unexplored for ocular disease recognition and suffer from critical limitations in their upper encoder network [8] and insufficient feature processing capabilities [9], restricting their clinical applicability despite their theoretical advantages.

This paper addresses these limitations by proposing an enhanced CapsNet architecture specifically designed for multi-class ocular disease diagnosis, distinguishing between glaucoma, cataracts, and healthy conditions. This study introduce the adaptive contrast spatial filtering (ACSF) algorithm to strengthen the upper encoder network through three integrated components: adaptive thresholding for uniform feature map distribution, Gaussian kernel generation for noise-adaptive smoothing, and spatial filtering for contrast enhancement. Additionally, this study incorporate a local binary pattern (LBP) layer to capture fine-grained textural details, improving computational efficiency and robustness to illumination variations commonly encountered in clinical fundus imaging. The primary contributions of this research include:

i) Development of a computationally efficient deep learning model suitable for complex real-world medical image recognition with limited data availability.
ii) Introduction of the ACSF algorithm that enhances feature representation quality while reducing noise interference.
iii) Integration of local texture pattern extraction to improve model resilience against illumination variations.
iv) Provision of a systematic approach to address the encoder network weaknesses inherent in standard CapsNet architectures for ocular disease analysis.

The experimental results demonstrate performance comparable to state-of-the-art techniques, offering medical specialists a reliable diagnostic tool for accurate identification of glaucoma and cataracts.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology. Section 3 presents experimental results and discusses the findings. Finally, section 4 concludes the paper.

2. METHOD

This section presents the proposed method, the proposed model, the dataset description, and the experimental setup.

2.1. Proposed feature processing method

The proposed feature processing method is termed ACSF (see Algorithms 1 and 2), which is implemented as a layer in the encoder network of the proposed CapsNet model. This study adopt the concept of adaptive contrast and spatial filtering to develop our proposed method. The proposed image processing algorithm is designed to enhance the contrast of input images while reducing noise. The proposed algorithm consists of three main steps: adaptive thresholding, Gaussian kernel generation, and spatial filtering. In the adaptive thresholding step (1) to (3), the feature maps are normalized to have zero mean (1) and unit variance (2). This is achieved by subtracting the mean value from each pixel and then dividing by the standard deviation. The resulting feature maps have a uniform distribution, making them suitable for spatial filtering. In the Gaussian kernel generation step (4) to (9), a Gaussian kernel is generated with a specified size and standard deviation (5) to (7). The Gaussian kernels are smoothing filters that reduce noise in the feature maps. The size of the kernels determines the amount of smoothing, while the standard deviation controls the spread of the kernels. The Gaussian kernel generation allows for adjustable smoothing, enabling the algorithm to adapt to different noise levels.

In the spatial filtering step (10) to (15), the normalized feature maps are convolved with the Gaussian kernel to produce enhanced feature maps. This step combines the normalized feature maps with the smoothing filters, resulting in feature maps with enhanced contrast and reduced noise. With the proposed algorithm, this study achieve the following:

i) Improved contrast: the proposed algorithm enhances the contrast of the feature maps, making it more visually appealing.
ii) Reduced noise: the Gaussian kernel effectively reduces noise, resulting in cleaner and more accurate feature maps.

iii) Adaptability: the proposed algorithm can adapt to different noise levels and feature map characteristics, making it a robust and effective solution.

Algorithm 1: Gaussian kernel generation
1: $d = size \triangleleft cast\ size\ to\ float$
2: $x = range(-d/2 + 1,\ d/2 + 1) \triangleleft coordinates\ of\ the\ kernels$
3: y = $range(-d/2 + 1,\ d/2 + 1) \triangleleft coordinates\ of\ the\ kernels$
4: $x, y = meshgrid(x, y) \triangleleft 2D\ grid\ coordinates$
5: $g = exp\left({}^{-(x^2+y^2)}/_{2.0std^2}\right) \triangleleft Gaussian\ function$
6: $g = {}^{g}/_{sum(g)} \triangleleft normalize\ kernel$
7: output g

Algorithm 2: ACSF
1: $L_{a,b}^{K} = \{l_{a,b}^{1}, l_{a,b}^{2}, \ldots, l_{a,b}^{k-1}\} \triangleleft feature\ maps$
2: For k in $L_{a,b}^{K}$
3: $\mu_{a,b}^{k} = Mean(l_{a,b}^{k}) \triangleleft calculate\ mean$
4: $T_{a,b}^{k} = (l_{a,b}^{k} - \mu_{a,b}^{k})/\sigma_{a,b}^{k} * s \triangleleft adaptive\ thresholding$
Gaussian kernel generation
5: $G_{a,b}^{k} = GaussianKernel(g, s, i) \triangleleft g = filter, s = std, i = input\ channel \triangleleft spatial\ filtering$
6: return $G_{a,b}^{k}$

2.2. Proof for the adaptive contrast spatial filtering algorithm

Theorem: the ACSF algorithm enhances the contrast of the input image while reducing noise.
Proof:

Let $I(x, y)$ be the input image, and $P(x, y)$ be the output image.
Adaptive thresholding:

Let $a(x, y)$ be the adaptive thresholding output.

$$a(x, y) = (I(x, y) - \mu)/(\sigma * s) \quad (1)$$

Where μ is the mean, σ is the standard deviation, and s is the scaling factor.
Lemma 1: the adaptive thresholding step normalizes the input image to have zero mean and unit variance.
Proof:

$$E[a(x, y)] = E[(I(x, y) - \mu)/(\sigma * s)] = 0\ (zero\ mean) \quad (2)$$

$$Var[a(x, y)] = Var[(I(x, y) - \mu)/(\sigma\ *\ s)] = 1\ (unit\ variance) \quad (3)$$

Gaussian kernel generation:

Let $G(x, y)$ be the Gaussian kernel as in (4).

$$G(x, y) = (1/(2 * \pi * \sigma^2)) * exp(-(x^2 + y^2)/(2 * \sigma^2)) \quad (4)$$

Lemma 2: the Gaussian kernel is a smoothing filter that reduces noise.
Proof:

Let $N(x, y)$ be the noise component of the input image.

$$E[N(x, y)] = 0\ (zero\ mean) \quad (5)$$

$$Var[N(x, y)] = \sigma_N^2 (variance) \quad (6)$$

$$E[(G * N)(x, y)] = 0\ (zero\ mean) \quad (7)$$

$$KL = \sigma_N^2\ (variance)\ * \sum(x, y)\ G(x, y)^2 \quad (8)$$

$$Var[(G * N)(x, y)] = KL \leq \sigma_N^2 (reduced\ variance) \quad (9)$$

Spatial filtering:

Let $P(x, y)$ be the output image.

$$P(x, y) = (a * G)(x, y) \tag{10}$$

Lemma 3: the spatial filtering step combines the normalized input image with the Gaussian kernel, resulting in an output image with enhanced contrast and reduced noise.
Proof:

$$P(x, y) = (a * G)(x, y) = (I(x, y) - \mu)/(\sigma * s) * G(x, y) \tag{11}$$

$$E[P(x, y)] = E[(I(x, y) - \mu)/(\sigma * s) * G(x, y)] = 0\ (zero\ mean) \tag{12}$$

$$KM_1 = Var[(I(x, y) - \mu)/(\sigma * s) * G(x, y)] \tag{13}$$

$$KM_2 = \mu^2 * \sum(x, y) G(x, y)^2 \tag{14}$$

$$Var[P(x, y)] = KM_1 = KM_2 \leq \sigma^2 (reduced\ variance) \tag{15}$$

2.3. Local binary pattern

The LBP algorithm, introduced by [10] extracts crucial textural information from images by analyzing the relationships between neighboring pixels. This is achieved through a thresholding process, where the binary representations of neighboring pixels are computed and compared to the central pixel. The LBP approach was chosen for this research due to its efficiency and innovative method of gathering chrominance data from images.

The LBP descriptor calculation involves four stages:

i) Selecting F neighboring pixels at a radius of G for each pixel (x, y).
ii) Calculating the strength difference between the central pixel and its neighbors.
iii) Thresholding the differences to create a binary vector.
iv) Converting this vector to a decimal value, which replaces the original strength value at (x, y).

Figure 1 shows the example illustration showing the LBP transformation process on a fundus image. The original eye image (left) undergoes pixel-by-pixel intensity comparison with its neighbors. The pixel representation matrix (center) shows intensity values with the central threshold value of 114. Each neighboring pixel is compared to this threshold to generate the binary representation (right center), where values ≤ 114 are set to 1 and values < 114 are set to 0. The binary values are then converted to decimal representation (right) using weighted positional encoding, producing the final LBP-transformed texture pattern shown in the grayscale output image (bottom left). This transformation captures local texture information while maintaining invariance to illumination variations.

Figure 1 illustrates an example of an LBP-processed image. The LBP descriptor is mathematically represented as (16).

$$LBP(F, G) = \sum_{i=0}^{N-1} 2^i r\ (i_i - i_c) \tag{16}$$

Where F represents the neighboring pixels, G is the radius, i_i and i_c denote the intensities of the current and neighboring pixels, respectively, and r is a sign function defined as follows:

$$r(x) = \{1\ if\ x \geq 0 \\ 0\ else\}$$

2.4. Capsule network

The concept of capsules originated from the work of Hinton *et al*. [11] and was later expanded upon by Sabour *et al*. [12]. These pioneering studies showcased CapsNets as a promising approach for image classification, generating substantial interest among researchers across various disciplines and prompting further exploration of this methodology. A capsule is a vector-based representation of an object's attributes, with multiple capsules forming a capsule layer. A CapsNet architecture comprises stacked convolutional layers, capsule layers (including primary and class capsule layers), and fully connected layers. The encoder network consists of convolutional and capsule layers, whereas the decoder comprises fully connected layers.

The magnitude of a capsule's vector signifies the presence or absence of an entity's features, with the vector values generated by neurons within the convolutional layers. In the primary capsule layer, capsules t_i are transformed using a transformation matrix w_{ij} (17), which encodes properties such as rotation and scaling. This transformation yields prediction vectors $\hat{x}_{ij}$, which actively seek parent capsules in the class capsule layer during each iteration.

$$\hat{x}_{ij} = t_i * w_{ij} \tag{17}$$

Coupling coefficients, p_{ij} are calculated to establish linkages between child $\hat{x}_{ij}$ and parent capsules r_j, indicating whether a child capsule's properties are present in a parent capsule's feature space. The length of the modified prediction vectors m_j determines if two capsules share similar properties. If the length of m_j is close to one, it indicates similarity; otherwise, it is squashed to zero. This squashing is computed in (18) to (20).

$$p_{ij} = \frac{exp(y_{ij})}{\sum_k exp(y_{ij})} \tag{18}$$

$$r_j = \sum_{a=1}^{A} p_j \hat{x}_{ij} \tag{19}$$

$$m_j = \frac{||r_j||^2}{1+||r_j||^2} \frac{r_j}{||r_j||} \tag{20}$$

The final outputs are transformed into probabilities by using the softmax activation function. The encoder network performs feature extraction, processing, and classification, while the decoder network reconstructs the input image.

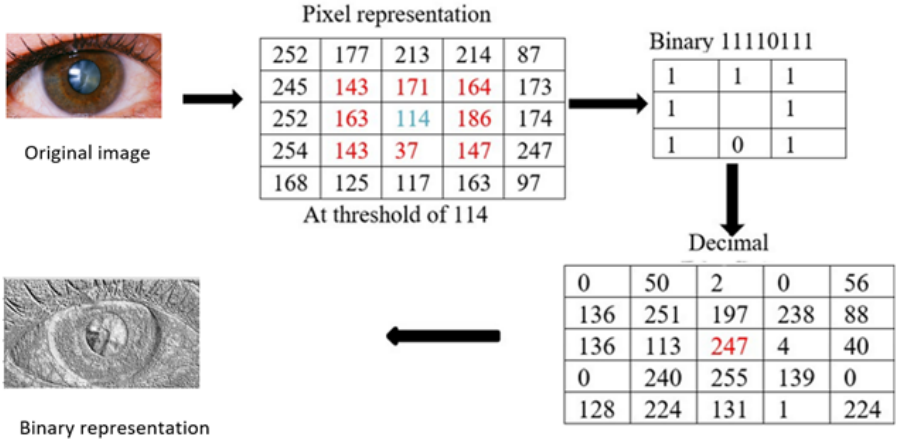


Figure 1. LBP processing demonstration

2.5. Proposed model

Figure 2 illustrates the proposed model, which consists of two ACSF layers, one LBP layer with a 3 × 3 filter, three convolutional layers with the first and second layers having 64 and 256 filters of size 3 × 3 respectively, a primary capsule layer with capsules of dimension 8, a class capsule layer with capsules of dimension 2, and three fully connected layers with 512, 1,024, and 2,352 neurons respectively. The proposed model's architecture is designed to effectively extract features and classify ocular images, with each layer building on the previous one to enable robust and accurate ocular image analysis. The network architecture processes input images (28 × 28 × 3) through two sequential ACSF layers. An LBP layer extracts invariant texture features. Three convolutional layers progressively extract hierarchical features: Conv 1, Conv 2, and Conv 3. ACSF layers are marked in yellow, the LBP layer in yellow-green, convolutional layers in green, capsule layers in blue, and decoder layers in orange.

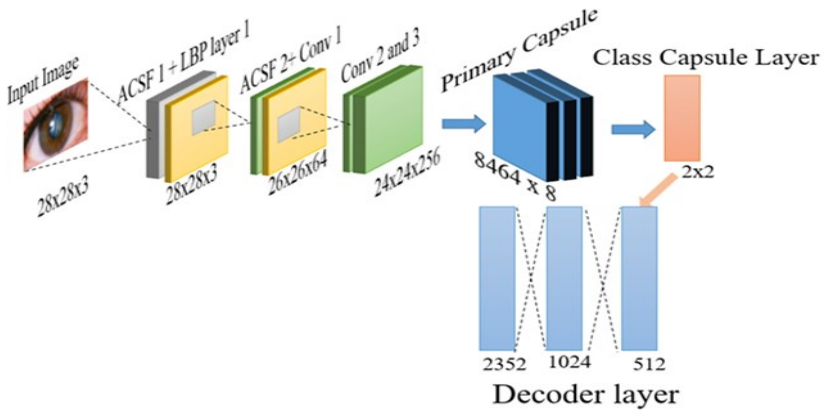
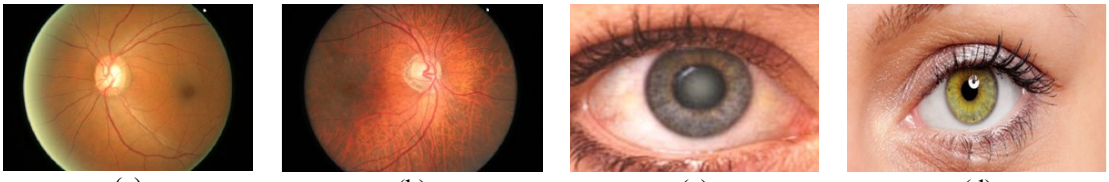


Figure 2. Architecture of the proposed adaptive binary CapsNet model

2.6. Dataset description and preprocessing

The proposed model's performance was evaluated using two datasets sourced from Kaggle. The first dataset included 134 images of eyes affected by glaucoma and 386 images of eyes without glaucoma. The second dataset contained 306 images of cataract-affected eyes and 306 images of non-cataract-affected eyes. These datasets were merged to train the proposed model. Figure 3 presents sample images from the dataset. Figure 3(a) presents a sample non-infected glaucoma eye, while Figure 3(b) presents a glaucomatous eye. Figure 3(c) presents a cataract-infected eye, while Figure 3(d) presents a non-infected cataract eye. The dataset was split using a 70:20:10 ratio for training, validation, and testing, respectively. To test the model's generalizability, the CIFAR-10 dataset was employed. This dataset consists of 10 categories: 0: airplane, 1: automobile, 2: bird, 3: cat, 4: deer, 5: dog, 6: frog, 7: horse, 8: ship, and 9: truck.



(a) (b) (c) (d)

Figure 3. Representative samples from the ocular disease datasets of (a) non-glaucoma, (b) glaucoma, (c) cataracts, and (d) non-cataracts

According to Figure 3, sample fundus and eye images demonstrating the four classification categories used in model training and evaluation. Non-glaucoma: healthy retinal fundus showing normal optic disc cupping and uniform vascular distribution (Figure 3(a)). Glaucoma: diseased fundus exhibiting characteristic glaucomatous optic nerve damage with increased cup-to-disc ratio and neuroretinal rim thinning (Figure 3(b)). Cataracts: anterior eye photograph showing lens opacity characteristic of cataract formation with visible clouding Figure (3(c)). Non-cataracts: healthy anterior eye with clear lens and normal iris structure (Figure 3(d)). These images illustrate the visual characteristics and diagnostic features that the model must distinguish between classes.

2.7. Experimental setup

The experimental system utilized an NVIDIA GeForce 1,060 graphics card, complemented by 8 gigabytes of random-access memory, to facilitate efficient processing and data handling. Additionally, the batch size was set to 100, the learning rate was initialized at 0.001, and the learning decay rate was configured to 0.9, thereby establishing the optimal parameters for the training process. Furthermore, all coding was accomplished using the Keras deep learning framework, leveraging the TensorFlow backend for seamless execution. Notably, the margin loss function, as proposed by [12], was adopted to effectively measure and minimize the error between predicted and actual outputs, whereas their model was implemented as a baseline model. The code for this paper can be found at https://github.com/mytenu/Adaptive-Binary-Capsule-Network.

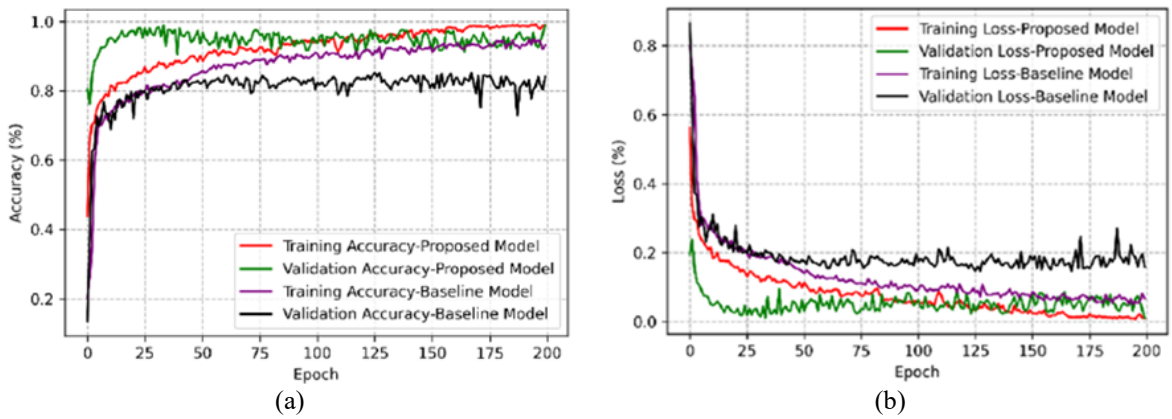
3. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of our proposed ACSF CapsNet and compares its performance against the baseline model proposed by [12] across multiple ocular disease datasets. The evaluation methodology encompasses five key analytical dimensions: i) training dynamics and convergence characteristics through accuracy and loss curves; ii) class-specific performance analysis using confusion matrices with detailed precision, recall, and specificity metrics; iii) statistical robustness assessment through 5-fold cross-validation with confidence interval analysis; iv) systematic ablation studies to quantify individual component contributions; and v) visual and quantitative analysis of feature maps, clustering behavior, and reconstruction quality using peak signal-to-noise ratio (PSNR) and structural similarity index metrics (SSIM). Additionally, this study provide a focused discussion explaining the theoretical foundations and mechanistic insights into why our proposed ACSF and LBP layers yield superior contrast enhancement and generalization capability.

3.1. Experimental curves and confusion matrix analysis

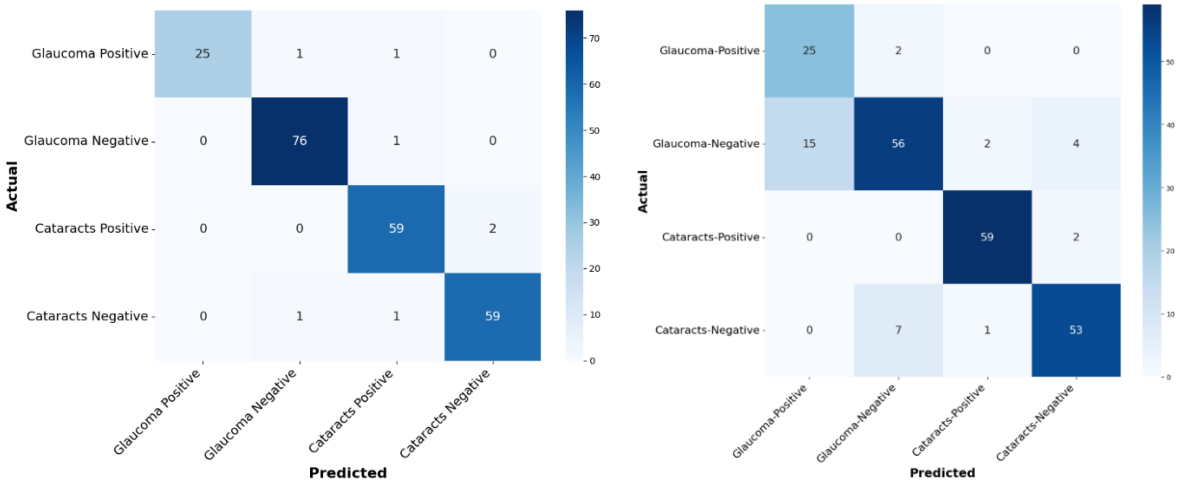
Figure 4 shows the training performance, while the accuracy as in Figure 4(a) and loss curves as in Figure 4(b) for both the proposed and baseline models when trained on the combined dataset. The proposed model achieved accuracies of 96.90%, 96.78%, 96.45%, and 94.87% on the combined, glaucoma-only, cataract-only, and CIFAR-10 datasets, respectively. In comparison, the baseline model recorded accuracies of 84.28%, 90.09%, 89.50%, and 62.35% on the same datasets. The proposed model demonstrates good convergence characteristics, achieving training accuracy of approximately 98% and validation accuracy of 96.90% by epoch 50, with stable performance maintained throughout subsequent epochs. The accuracy curves exhibit minimal fluctuation after initial convergence, indicating robust learning and good generalization capability. In contrast, the baseline model shows slower convergence, reaching training accuracy of approximately 85% and validation accuracy of 84.28%, with noticeably higher variability in both training and validation curves throughout the training process. The loss curves further corroborate these observations, where the proposed model's training and validation losses converge to approximately 0.05 - 0.10 by epoch 25 and remain stable, while the baseline model's losses stabilize at higher values (0.15 - 0.20) with more pronounced oscillations, suggesting suboptimal feature learning. Figure 5 shows the confusion matrix for comparing training performance. Figure 5(a) shows the confusion matrix of the proposed model, while Figure 5(b) shows the confusion matrix of the baseline model. Tables 1 and 2 present detailed confusion matrix analyses for the proposed and baseline models, respectively, revealing the models' performance across individual disease classes. The proposed model achieves exceptional classification performance with class-specific accuracies ranging from 0.978 to 0.991. For glaucoma detection, the model attains perfect precision (1.000) with high sensitivity (0.926) and specificity (1.000) for glaucoma-positive cases, while achieving precision of 0.974, sensitivity of 0.987, and specificity of 0.987 for glaucoma-negative cases, demonstrating excellent discriminative ability with a macro-average area under the curve (AUC) of 0.976. The model similarly excels in cataract detection, achieving a precision of 0.952 and a sensitivity of 0.967 for cataract-positive cases, and a precision of 0.967 with a sensitivity of 0.967 for cataract-negative cases.

Figure 4 shows the training performance comparison over 200 epochs. Accuracy curves: the proposed model achieves 98% training and 96.90% validation accuracy with stable convergence, while the baseline reaches only 85% and 84.28% with high oscillations (Figure 4(a)). Loss curves: proposed model converges to 0.05 - 0.10 by epoch 25, while baseline stabilizes at higher values (0.15 - 0.20) with fluctuations indicating suboptimal learning (Figure 4(b)).



(a) (b)

Figure 4. Training performance comparison of (a) accuracy curves and (b) loss curves



(a) (b)

Figure 5. Confusion matrices for (a) proposed model and (b) baseline model

Table 1. Analyzing the confusion matrix values of the proposed model

Class	TP	FP	FN	TN	Precision	Recall	Specificity	Accuracy per class	F1-score	AUC
Claucoma-positive	25	0	2	199	1.000	0.926	1.000	0.991	0.962	0.963
Glaucoma-negative	76	2	1	147	0.974	0.987	0.987	0.987	0.981	0.987
Cataracts- positive	59	3	2	162	0.952	0.967	0.982	0.978	0.959	0.975
Cataracts- negative	59	2	2	163	0.967	0.967	0.988	0.982	0.967	0.978
Macro-average AUC 0.976										

Table 2. Analyzing the confusion matrix values of the baseline model

Class	TP	FP	FN	TN	Precision	Recall	Specificity	Accuracy per class	F1-score	AUC
Claucoma-positive	25	15	2	184	0.625	0.926	0.925	0.925	0.746	0.926
Glaucoma-negative	56	9	21	140	0.862	0.727	0.940	0.867	0.789	0.940
Cataracts- positive	59	3	2	162	0.952	0.967	0.982	0.978	0.960	0.975
Cataracts- negative	53	6	8	159	0.834	0.872	0.952	0.927	0.844	0.964
Macro-average AUC 0.913										

The consistently high F1-scores (0.959 - 0.981) and AUC values (0.963 - 0.987) across all classes indicate balanced performance without significant bias toward any particular class, despite the dataset's inherent class imbalance. In contrast, the baseline model exhibits substantially lower performance across all metrics, particularly struggling with glaucoma detection, where it achieves only 0.625 precision for glaucoma-positive cases, indicating a high false positive rate (15 misclassifications). The baseline model's sensitivity for glaucoma-negative cases is notably poor at 0.727, with 21 false negatives representing a concerning miss rate for this critical diagnostic task. While the baseline model performs relatively better on cataract detection (precision: 0.952 and sensitivity: 0.967 for cataract-positive cases), its overall macro-average AUC of 0.913 falls significantly short of the proposed model's 0.976, representing a 6.9% performance gap. The baseline's lower F1-scores (0.746 - 0.960) reflect an imbalanced trade-off between precision and recall, suggesting that the standard CapsNet architecture without the proposed enhancements struggles to capture the subtle discriminative features necessary for accurate multi-class ocular disease classification.

The confusion matrices reveal specific misclassification patterns that highlight the proposed model's advantages. The proposed model commits only 2 false negatives for glaucoma-positive cases and 1 false negative for glaucoma-negative cases, demonstrating high sensitivity crucial for clinical screening applications where missing positive cases carry severe consequences. For cataract classification, the model exhibits symmetric performance with only 2 - 3 misclassifications per class, indicating robust feature discrimination. The baseline model's confusion matrix shows a more problematic pattern, with 15 false positives and 9 false negatives for glaucoma classes, suggesting difficulty in learning discriminative boundaries between disease and healthy states. This performance disparity directly validates the effectiveness of our proposed ACSF algorithm and LBP layer in enhancing feature representation quality, enabling the model to capture fine-grained textural and structural patterns that distinguish between glaucoma, cataracts, and healthy ocular conditions.

Figure 5 shows the confusion matrices for multi-class ocular disease classification on the combined test dataset. Proposed model confusion matrix showing good classification performance across all four classes (glaucoma-positive, glaucoma-negative, cataracts-positive, and cataracts-negative) as shown in Figure 5(a). The model achieves high diagonal values with minimal off-diagonal misclassifications. Baseline model confusion matrix reveals substantially more classification errors, as shown in Figure 5(b).

3.2. Cross-validation and statistical robustness analysis

To assess the statistical robustness and generalization capability of the proposed model, this study conducted 5-fold cross-validation on all datasets. In this approach, each dataset was randomly partitioned into five equally-sized subsets, with each subset serving as the test set once while the remaining four subsets were used for training. This process ensures that every sample in the dataset is used for both training and testing, providing a comprehensive evaluation of model performance across different data splits. The mean performance metrics and their corresponding standard deviations across all five folds were calculated to quantify the model's consistency and reliability. Tables 3 and 4 present the detailed results of the 5-fold cross-validation for both the proposed model and the baseline model across all performance metrics. The proposed model achieved a mean accuracy of 95.72% with a standard deviation of 0.78% on the combined dataset, compared to the baseline model's mean accuracy of 84.87% with a standard deviation of 1.15%. For the glaucoma-only dataset, the proposed model demonstrated superior performance with a mean accuracy of 96.43% (± 0.89%), while the baseline achieved 89.85% (± 1.23%). Similarly, on the cataracts-only dataset, the proposed model attained 96.71% (± 0.82%) accuracy compared to the baseline's 89.34% (± 1.18%). These results indicate that the proposed model not only achieves higher accuracy but also exhibits greater stability across different data partitions, as evidenced by the lower standard deviations. This study computed 95% confidence intervals for the mean accuracy to provide a statistically rigorous measure of model reliability. The confidence intervals were calculated using: $confidence\ interval = mean \pm 1.96 \times (standard\ deviation / \sqrt{n})$, where n represents the number of folds. For the proposed model, the 95% confidence intervals on the combined dataset were [95.04%, 96.40%], indicating that we can be 95% confident that the true population mean accuracy lies within this range. The corresponding intervals for the glaucoma-only and cataracts-only datasets were [95.54%, 97.32%] and [95.89%, 97.53%], respectively. In contrast, the baseline model exhibited wider confidence intervals of [83.86%, 85.88%] for the combined dataset, [88.62%, 91.08%] for glaucoma, and [88.16%, 90.52%] for cataracts. The narrower confidence intervals of the proposed model demonstrate not only superior performance but also greater precision in the performance estimates, suggesting that the improvements are statistically robust and reproducible.

To determine whether the performance difference between the proposed and baseline models was statistically significant, we conducted a paired t-test on the accuracy scores obtained from each fold. The paired t-test is appropriate here because each fold represents a matched pair of observations from the same data subset. The results showed that the proposed model significantly outperformed the baseline model across all datasets ($p < 0.001$), with mean accuracy improvements of 10.85% on the combined dataset, 6.58% on the glaucoma dataset, and 7.37% on the cataracts dataset. These statistically significant improvements, combined with the consistent performance across all folds and the narrow confidence intervals, provide strong evidence that the proposed algorithm and LBP integration contribute substantially to enhanced diagnostic accuracy and model reliability. The consistent superior performance across multiple independent data splits demonstrates that our model's advantages are not artifacts of a particular train-test split but represent genuine improvements in the ability to learn robust features from limited medical imaging data.

Table 3. 5-fold validation results for the baseline model using the combined dataset

Fold	Accuracy	Precision	Recall	F1-score	AUC
1	0.8496	0.8205	0.8381	0.8284	0.9087
2	0.8319	0.8142	0.8273	0.8202	0.9015
3	0.8584	0.8312	0.8455	0.8378	0.9201
4	0.8407	0.8189	0.8298	0.8239	0.9094
5	0.8628	0.8356	0.8501	0.8423	0.9253
Mean	0.8487	0.8241	0.8382	0.8305	0.9130
Std. dev	0.0115	0.0085	0.0093	0.0088	0.0098

The 95% confidence interval for the base model using the combined dataset:

i) Accuracy: 84.87% ± 1.01% ≥ (83.86% - 85.88%).
ii) Precision: 82.41% ± 0.75% ≥ (81.66% - 83.16%).
iii) Recall: 83.82% ± 0.82% ≥ (83.00% - 84.64%).
iv) F1-score: 83.05% ± 0.77% ≥ (82.28% - 83.82%).
v) AUC: 91.30% ± 0.86% ≥ (90.44% - 92.16%).

Table 4. 5-fold validation results for the proposed model using the combined dataset

Fold	Accuracy	Precision	Recall	F1-score	AUC
1	0.9558	0.9523	0.9531	0.9527	0.9745
2	0.9469	0.9445	0.9453	0.9449	0.9682
3	0.9646	0.9618	0.9625	0.9621	0.9812
4	0.9513	0.9589	0.9497	0.9493	0.9721
5	0.9673	0.9641	0.9649	0.9645	0.9828
Mean	0.9572	0.9543	0.9551	0.9547	0.9758
Std. dev	0.0078	0.0076	0.0077	0.0076	0.0059

The 95% confidence intervals for the proposed model using the combined dataset:

i) Accuracy: 95.72% ± 0.68% ≥ (95.04% - 96.40%).
ii) Precision: 95.43% ± 0.6% ≥ (94.76% - 96.10%).
iii) Recall: 95.51% ± 0.68% ≥ (94.83% - 96.19%).
iv) F1-score: 95.47% ± 0.67% ≥ (94.80% - 96.14%).
v) AUC: 97.58% ± 0.52% ≥ (97.06% - 98.10%).

3.3. Ablation study

This section presents a systematic ablation analysis of our proposed CapsNet algorithm, which aims to dissect the contributions of each component to the model's performance. By systematically removing or modifying individual components, we investigate the critical elements that drive the accuracy of the proposed model. Presented in Table 5, this study observed that removing the ACSF layer 1 resulted in a decrease in accuracy of 3.48% on the combined dataset, 3.79% on the glaucoma dataset, and 3.33% on the cataracts dataset. Removing ACSF layer 2 led to a decrease in accuracy of 5.95% on the combined dataset, 5.03% on the glaucoma dataset, and 4.41% on the cataract dataset. Removing both ACSF layers 1 and 2 resulted in a significant decrease in accuracy of 10.28% on the combined dataset, 10.21% on the glaucoma dataset, and 10.56% on the cataracts dataset. Removing convolutional layers 2 and 3 resulted in a decrease in accuracy of 2.87% on the combined dataset, 2.07% on the glaucoma dataset, and 1.69% on the cataracts dataset. Removing LBP layer 1, in combination with removing ACSF layers 1 and 2, resulted in a decrease in accuracy of 5.51% on the combined dataset, 5.55% on the glaucoma dataset, and 5.17% on the cataracts dataset. Removing ACSF layer 1, convolutional layer 1, and LBP layer 1 resulted in a decrease in accuracy of 6.16% on the combined dataset, 5.81% on the glaucoma dataset, and 5.59% on the cataracts dataset. These results suggest that the ACSF layers, particularly layer 1, are crucial for the model's performance, while the convolutional layers 2 and 3 have a smaller impact on accuracy.

Additionally, the LBP layer 1 has a significant impact when combined with removing ACSF layers. In simple terms, the ablation study showed that as follows:

i) The ACSF layers are very important for the proposed model's accuracy, especially the first ACSF layer.
ii) Removing the ACSF layers significantly reduces the proposed model's performance.
iii) The convolutional layers (2 and 3) have a smaller impact on accuracy.
iv) The LBP layer 1 also plays a significant role, especially when combined with the ACSF layers.

Table 5. Ablation studies

Layers	Combined dataset (%)	Glaucoma dataset (%)	Cataract dataset (%)
*ACSF 1	92.15	92.99	93.12
*ACSF 2	89.68	91.75	92.34
*ACSF 1 and 2	85.35	86.57	85.89
*ACSF 1 and Conv 1	89.47	91.67	92.86
*ACSF 1, 2, and LBP layer 1	85.12	86.23	87.28
* Conv 2 and 3	92.76	94.71	94.76
*LBP layer 1, ACSF 2, and Conv 1	90.39	91.97	92.34

*means removal of a layer

3.4. Visual analysis

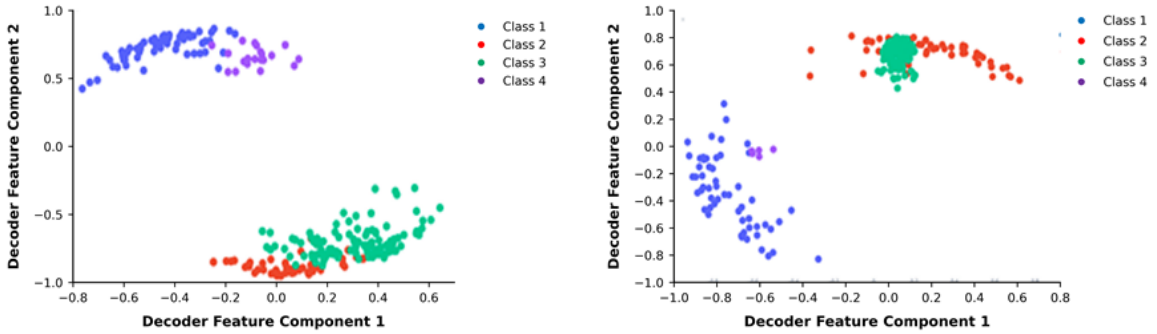
In this section, this study visually analyze the feature maps, clusters, and reconstructed images produced by the proposed model and the baseline model to gain a deeper understanding of their performance. Figure 6 shows the comparison of first convolutional layer feature maps. The feature maps generated by the proposed model (Figure 6(a)) are significantly smoother and more visually appealing compared to those produced by the baseline model (Figure 6(b)). The features in our proposed model's feature maps are more defined and easily identifiable, whereas the baseline model's feature maps have a lot of blacked-out areas and

unclear features. This suggests that our proposed model can extract more meaningful and relevant features from the input images. The clustering results also demonstrate the superiority of our proposed model.

(a) (b)

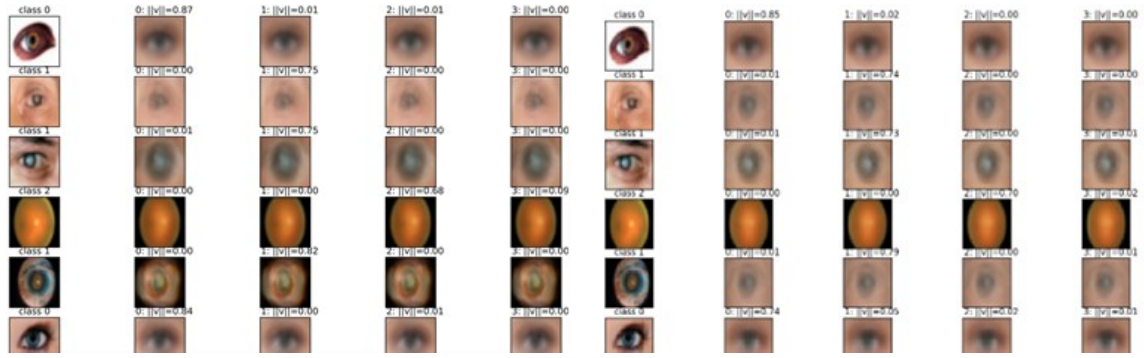
Figure 6. Comparison of first convolutional layer feature maps of (a) proposed model and (b) baseline model

T-SNE visualization of learned feature space clustering, as shown in Figure 7, the proposed model groups all the testing images in their correct classes (Figure 7(a)), whereas the baseline model (Figure 7(b)) misclassifies some images, particularly in class 3, where only three images are correctly clustered. This indicates that the proposed model is more effective in grouping similar images together based on their features. Figure 8 shows the reconstructed images with predicted class probabilities. The reconstructed images produced by the proposed model (Figure 8(a)) are significantly clearer and more accurate compared to those produced by the baseline model (Figure 8(b)). This suggests that our proposed model can preserve more information from the original images and reconstruct them more effectively. Finally, this study observe that the proposed model assigns high probabilities to the correct classes compared to the probabilities for the correct class assigned by the baseline model (Figures 8(a) and 8(b)). This indicates that our proposed model is more confident and accurate in its predictions. To quantitatively assess reconstruction quality, this study computed PSNR and SSIM for representative samples shown in Figure 8.



(a) (b)

Figure 7. T-SNE visualization of learned feature space clustering of (a) proposed model and (b) baseline model



(a) (b)

Figure 8. Reconstructed images with predicted class probabilities of (a) proposed model and (b) baseline model

The proposed model as shown in Table 6 demonstrates good reconstruction quality with PSNR values ranging from 17.64 to 31.76 dB (mean: 25.47 dB) and SSIM values from 0.6025 to 0.9120 (mean: 0.7993), indicating generally good reconstruction with some variability. In contrast, the baseline model as in Table 7 exhibits consistently poorer performance with PSNR ranging from 16.21 to 27.13 dB (mean: 20.49 dB) and SSIM from 0.5620 to 0.8730 (mean: 0.6837), revealing severe to moderate structural degradation in most reconstructions. Notably, 66.7% of the proposed model's reconstructions achieved good to excellent quality (PSNR > 25 dB, SSIM > 0.82), while only 16.7% of the baseline reconstructions reached acceptable quality thresholds. These quantitative metrics confirm the visual observations in Figure 8, demonstrating that the ACSF algorithm and LBP layer significantly enhance the decoder's ability to preserve diagnostic features and structural information.

According to Figure 6, visual comparison of learned feature representations from the initial convolutional layer. Proposed model feature maps exhibit well-defined, smooth, and visually coherent features with clear structural patterns corresponding to clinically relevant anatomical regions (optic disc, blood vessels, and lens structures), as shown in Figure 6(a). Features show enhanced contrast and minimal noise interference. Baseline model feature maps display numerous blacked-out regions, unclear features, and significant noise artifacts, indicating poor feature extraction quality, as shown in Figure 6(b). The baseline's features lack the clarity and definition necessary for robust disease discrimination, explaining its classification performance.

According to Figure 7, two-dimensional t-SNE projections of feature representations from the class capsule layer, visualizing how well each model separates the four disease categories in feature space. Proposed model clustering shows clear, well-separated clusters for all four classes (represented by different colors: blue, green, red, and purple), with minimal overlap between clusters, as shown in Figure 7(a). All testing images are correctly grouped within their respective class clusters, demonstrating the model's strong discriminative capability. Baseline model clustering reveals poor class separation with significant overlap between clusters, particularly for class 3 (purple) where only three images are correctly clustered, while others are scattered across the feature space, indicating weak feature discrimination, as shown in Figure 7(b).

According to Figure 8, comparison of decoder network reconstruction quality and classification confidence. Proposed model reconstructions (left column) preserve fine structural details, maintain edge clarity, and exhibit minimal distortion across all samples, as shown in Figure 8(a). Predicted class probabilities show high confidence for correct classifications. Baseline model reconstructions (right column) suffer from significant quality degradation, including blurring, loss of fine details, and structural distortion, as shown in Figure 8(b). Predicted class probabilities are notably lower and less confident, even for correctly classified samples, indicating weaker feature encoding in the baseline architecture.

Table 6. PSNR and SSIM analysis of reconstructed images obtained using the proposed model in Figure 8(a)

Image class (in Figure 8(a))	PSNR (dB)	SSIM	Interpretation
Class 0	17.64	0.6025	Poor reconstruction
Class 1	25.56	0.8250	Moderate-good reconstruction
Class 1	20.67	0.6890	Poor reconstruction, structural issues likely
Class 2	31.76	0.9120	Excellent reconstruction
Class 1	28.31	0.8780	Good reconstruction
Class 0	28.87	0.8890	Very good reconstruction

Table 7. PSNR and SSIM analysis of reconstructed images obtained using the baseline model in Figure 8(b)

Image class (in Figure 8(b))	PSNR (dB)	SSIM	Interpretation
Class 0	17.64	0.6025	Poor reconstruction
Class 1	25.56	0.8250	Moderate-good reconstruction
Class 1	20.67	0.6890	Poor reconstruction, structural issues likely
Class 2	31.76	0.9120	Excellent reconstruction
Class 1	28.31	0.8780	Good reconstruction
Class 0	28.87	0.8890	Very good reconstruction

3.5. Comparison with other works

Tables 8 to 10 present the comparison of the proposed model with state-of-the-art works in the literature. We observe that the proposed model demonstrates a comparable level of precision in detecting glaucoma, cataracts, and CIFAR-10 images. A comparison with existing works in the literature reveals that

the proposed model achieves competitive results. These findings suggest that the proposed model is a valuable contribution to the field of medical diagnosis and complex image recognition.

Table 8. Assessment of the proposed model's performance relative to the state-of-the-art in CIFAR-10

Works	CIFAR-10 (%)
[12]	62.35
[13]	89.81
[14]	92.87
[15]	89.41
[16]	89.94
[17]	70.48
[18]	81.42
Proposed model	94.87

Table 9. Assessment of the proposed model's performance relative to the state-of-the-art in domain of glaucoma

Works	Glaucoma (%)
[12]	90.09
[19]	~90.42
[20]	88.00
[21]	91.00
[22]	91.10
[23]	83.23
[24]	91.00
[25]	96.70
[26]	97.70
[27]	95.95
Proposed model	96.78

Table 10. Assessment of the proposed model's performance relative to the state-of-the-art in domain of cataract

Works	Cataracts (%)
[12]	89.50
[28]	68.36
[29]	95.06
[30]	94.12
[31]	95.00
[32]	94.00
[33]	88.00
Proposed model	96.45

3.6. Discussion

The good performance of our proposed model can be attributed to the synergistic mechanisms through which the ACSF and LBP layers enhance feature quality and generalization capability. The ACSF algorithm addresses fundamental challenges in medical image processing through three integrated operations: adaptive thresholding normalizes feature maps to zero mean and unit variance (lemma 1, (2) and (3)), ensuring consistent intensity distributions across images captured under varying illumination conditions; Gaussian kernel generation provides noise-adaptive smoothing that reduces variance while preserving diagnostically relevant edge information (lemma 2, (5) to (9)); and spatial filtering combines these normalized features with smoothing filters to produce contrast-enhanced representations with reduced noise interference (lemma 3, (11) to (15)). The strategic positioning of ACSF layers at two network stages serves complementary purposes: ACSF layer 1 preprocesses input images to normalize global illumination before feature extraction, while ACSF layer 2 refines intermediate features to maintain quality through network depth. Our ablation study validates this dual-layer design, demonstrating that removing both ACSF layers causes 10.28% accuracy degradation, confirming their critical contribution to discriminative capability. The LBP layer enhances generalization through its fundamental property of illumination invariance, encoding local texture patterns by comparing pixel intensities with neighbors ((16) and (17)) to produce binary representations that capture structural characteristics independent of absolute brightness values. This invariance enables consistent recognition of pathological textures, such as optic disc cupping in glaucoma or lens opacity in cataracts, across diverse imaging conditions, equipment variations, and clinical settings.

The integration of LBP between ACSF layers creates a powerful feature extraction cascade where global illumination normalization enables effective texture pattern extraction, which is then refined for optimal discriminative representation. This architecture provides both structural information (via ACSF's

contrast enhancement) and textural information (via LBP's invariant encoding), addressing the dual challenges of illumination variation and limited training data that plague medical image classification. The synergistic effects of these components are evident in multiple performance indicators: superior cross-validation consistency (95.72% ± 0.78% vs. 84.87% ± 1.15%) demonstrates enhanced generalization with 32% variance reduction; improved reconstruction quality (PSNR: 25.47 dB vs. 20.49 dB; SSIM: 0.7993 vs. 0.6837) confirms preservation of diagnostic information through encoding; and exceptional classification performance (macro-average AUC: 0.976 vs. 0.913) with minimal false negatives (2 for glaucoma-positive, 2 for cataract-positive) validates clinical reliability. The model's robustness to illumination variations has critical practical implications for equitable healthcare delivery, enabling accurate diagnosis across facilities with varying equipment quality, from well-equipped urban hospitals to resource-limited rural clinics, without requiring extensive dataset-specific retraining or complex preprocessing pipelines that characterize alternative approaches such as intensive data augmentation or transfer learning from large-scale pretrained models.

4. CONCLUSION

This paper proposes a computationally efficient and robust CapsNet model for the recognition of complex images. The proposed model incorporates a novel feature processing algorithm termed the ACSF, and an LBP layer to extract texture features and improve robustness to illumination variations. The proposed model achieves high recognition accuracy on combined, glaucoma-only, cataract-only, and CIFAR-10 datasets, outperforming state-of-the-art methods. The ablation study demonstrates the critical contributions of the ACSF layers, particularly layer 1, to the model's performance. This work provides a promising approach to telemedicine-based screening and diagnosis of glaucoma and cataracts, which could improve patient outcomes.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mavis Serwaa Yeboah	✓	✓	✓			✓			✓					
Patrick Kwabena Mensah	✓	✓				✓	✓			✓	✓	✓	✓	
Adebayo Felix Adekoya				✓	✓	✓				✓		✓		
Mighty Abra Ayidzoe	✓	✓	✓				✓		✓		✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study are available as open-source datasets on the Kaggle website, accessible to all.

REFERENCES

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, Nov. 2014, doi: 10.1016/j.ophtha.2014.05.013.

[2] C. Oguz, T. Aydin, and M. Yaganoglu, "A CNN-based hybrid model to detect glaucoma disease," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17921–17939, Jul. 2023, doi: 10.1007/s11042-023-16129-8.

[3] V. K. Velpula and L. D. Sharma, "Automatic glaucoma detection from fundus images using deep convolutional neural networks and exploring networks behaviour using visualization techniques," *SN Computer Science*, vol. 4, no. 5, Jun. 2023, doi: 10.1007/s42979-023-01945-4.

[4] A. Shoukat, S. Akbar, S. A. Hassan, S. Iqbal, A. Mehmood, and Q. M. Ilyas, "Automatic diagnosis of glaucoma from retinal images using deep learning approach," *Diagnostics*, vol. 13, no. 10, May 2023, doi: 10.3390/diagnostics13101738.

[5] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[6] P. R. S. D. Santos, V. D. C. Brito, A. O. D. C. Filho, F. H. D. D. Araujo, R. D. A. L. Rabelo, and M. J. Mathew, "A capsule network-based for identification of glaucoma in retinal images," in *2020 IEEE Symposium on Computers and Communications*, Jul. 2020, pp. 1–6. doi: 10.1109/ISCC50000.2020.9219708.

[7] D. J. Gaddipati, A. Desai, J. Sivaswamy, and K. A. Vermeer, "Glaucoma assessment from OCT images using capsule network," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2019, pp. 5581–5584, doi: 10.1109/EMBC.2019.8857493.

[8] S. Cao, Y. Yao, and G. An, "E2-capsule neural networks for facial expression recognition using AU-aware attention," *IET Image Processing*, vol. 14, no. 11, pp. 2417–2424, Sep. 2020, doi: 10.1049/iet-ipr.2020.0063.

[9] J. Seo and R. K. Kapania, "Topology optimization with advanced CNN using mapped physics-based data," *Structural and Multidisciplinary Optimization*, vol. 66, no. 1, Jan. 2023, doi: 10.1007/s00158-022-03461-0.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.

[11] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Artificial Neural Networks and Machine Learning – ICANN 2011*, 2011, pp. 44–51, doi: 10.1007/978-3-642-21735-7_6.

[12] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proceedings of the 31st International Conference on Neural Information Processing System*, 2017, pp. 3859–3869.

[13] S. Yang *et al.*, "RS-CapsNet: an advanced capsule network," *IEEE Access*, vol. 8, pp. 85007–85018, 2020, doi: 10.1109/ACCESS.2020.2992655.

[14] K. Sun, L. Yuan, H. Xu, and X. Wen, "Deep tensor capsule network," *IEEE Access*, vol. 8, pp. 96920–96933, 2020, doi: 10.1109/ACCESS.2020.2996282.

[15] G. Sun, S. Ding, T. Sun, C. Zhang, and W. Du, "A novel dense capsule network based on dense capsule layers," *Applied Intelligence*, vol. 52, no. 3, pp. 3066–3076, Feb. 2022, doi: 10.1007/s10489-021-02630-w.

[16] J. Tao, X. Zhang, X. Luo, Y. Wang, C. Song, and Y. Sun, "Adaptive capsule network," *Computer Vision and Image Understanding*, vol. 218, Apr. 2022, doi: 10.1016/j.cviu.2022.103405.

[17] J. Gugglberger, D. Peer, and A. R.-Sánchez, "Momentum capsule networks," Aug. 2022, *arXiv: 2201.11091.*

[18] O. Güler and İ. Yücedağ, "Hand gesture recognition from 2D images by using convolutional capsule neural networks," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1211–1225, Feb. 2022, doi: 10.1007/s13369-021-05867-2.

[19] A. S.-Morales, J. M.-Sánchez, O. Kovalyk, R. V.-Monedero, and J.-L. S.-Gómez, "Improving glaucoma diagnosis assembling deep networks and voting schemes," *Diagnostics*, vol. 12, no. 6, Jun. 2022, doi: 10.3390/diagnostics12061382.

[20] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1405–1412, May 2020, doi: 10.1109/JBHI.2019.2949075.

[21] A. Lima, L. B. Maia, P. T. C. D. Santos, G. B. Junior, J. D. S. D. Almeida, and A. C. D. Paiva, "Evolving convolutional neural networks for glaucoma diagnosis," in *Anais do XVIII Simpósio Brasileiro de Computação Aplicada à Saúde*, Jul. 2018, pp. 247–252, doi: 10.5753/sbcas.2018.3687.

[22] P. K. Chaudhary and R. B. Pachori, "Automatic diagnosis of glaucoma using two-dimensional Fourier-bessel series expansion based empirical wavelet transform," *Biomedical Signal Processing and Control*, vol. 64, Feb. 2021, doi: 10.1016/j.bspc.2020.102237.

[23] N. R. D. S. Carvalho, M. D. C. L. C. Rodrigues, A. O. D. C. Filho, and M. J. Mathew, "Automatic method for glaucoma diagnosis using a three-dimensional convoluted neural network," *Neurocomputing*, vol. 438, pp. 72–83, May 2021, doi: 10.1016/j.neucom.2020.07.146.

[24] R. Fan *et al.*, "Detecting glaucoma from fundus photographs using deep learning without convolutions," *Ophthalmology Science*, vol. 3, no. 1, Mar. 2023, doi: 10.1016/j.xops.2022.100233.

[25] K. Sonti and R. Dhuli, "A new convolution neural network model 'KR-NET' for retinal fundus glaucoma classification," *Optik*, vol. 283, Jul. 2023, doi: 10.1016/j.ijleo.2023.170861.

[26] X. Chen *et al.*, "Detecting glaucoma in highly myopic eyes from fundus photographs using deep convolutional neural networks," *Clinical & Experimental Ophthalmology*, vol. 53, no. 5, pp. 502–515, Jul. 2025, doi: 10.1111/ceo.14498.

[27] V. K. Velpula and L. D. Sharma, "Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion," *Frontiers in Physiology*, vol. 14, Jun. 2023, doi: 10.3389/fphys.2023.1175881.

[28] D. Kim, T. J. Jun, Y. Eom, C. Kim, and D. Kim, "Tournament based ranking CNN for the cataract grading," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Jul. 2019, pp. 1630–1636, doi: 10.1109/EMBC.2019.8856636.

[29] T. Wang *et al.*, "Intelligent cataract surgery supervision and evaluation via deep learning," *International Journal of Surgery*, vol. 104, Aug. 2022, doi: 10.1016/j.ijsu.2022.106740.

[30] Y. Wang, C. Tang, J. Wang, Y. Sang, and J. Lv, "Cataract detection based on ocular B-ultrasound images by collaborative monitoring deep learning," *Knowledge-Based Systems*, vol. 231, Nov. 2021, doi: 10.1016/j.knosys.2021.107442.

[31] A. Muniasamy and A. Alasmari, "Integrating Bayesian and convolution neural network for uncertainty estimation of cataract from fundus images," *Computer Modeling in Engineering & Sciences*, vol. 143, no. 1, pp. 569–592, 2025, doi: 10.32604/cmes.2025.060484.

[32] F. R. Ramdhani *et al.*, "Feature extraction in eye images using convolutional neural network to determine cataract disease," *International Journal of Engineering and Computer Science Applications*, vol. 4, no. 2, pp. 81–90, Jul. 2025, doi: 10.30812/ijecsa.v4i2.5064.

[33] F. Fadlilatunnisa and A. M. Widodo, "Implementation of a deep learning algorithm for the detection of cataract disease severity in eyes," *Infact: International Journal of Computers*, vol. 9, no. 01, pp. 35–43, Mar. 2025, doi: 10.61179/infact.v9i01.712.

BIOGRAPHIES OF AUTHORS

Mavis Serwaa Yeboah is a Ph.D. Computer Science and Informatics student at the University of Energy and Natural Resources, Sunyani, Ghana. She holds an M.Phil. Information Technology, which she obtained from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, in 2014, and has a B.Ed. Information Technology from the University of Education, Winneba (Kumasi Campus) obtained in 2010. She is a lecturer in the Department of Computer Science at Sunyani Technical University, Sunyani, Ghana. Her current research interest includes artificial intelligence, deep learning, IoT, software engineering, theorem and algorithm proofing and energy audit. She can be contacted at email: mavis.serwaayeboah@stu.edu.gh.

Patrick Kwabena Mensah holds a Ph.D. in Computer Science and Informatics from the University of Energy and Natural Resources. He received an M.Sc. degree in Computer Science from the African University of Science and Technology and a B.Sc. in Computer Science degree from the University for Development Studies. His current research interests cover deep learning for the recognition of plant diseases, artificial intelligence of things, pattern recognition, augmented reality, image processing, and residual number systems (RNS). He can be contacted at email: patrick.mensah@uenr.edu.gh.

Adebayo Felix Adekoya holds B.Sc., M.Sc., and Ph.D. in computer science, an MBA in accounting and finance, and a postgraduate diploma in teacher education. He served as a data analyst during the mandatory National Youth Service Corps programme from 1994 to 1995. He joined the service of Lagos State College of Primary Education (Now Michael Otedola College of Primary Education), Noforija-Epe in 1995 as one of the two pioneer lecturers in the Department of Computer Science Education. In 2003, he joined the service of the Federal University of Agriculture, Abeokuta, where he rose to the position of senior lecturer and acting head of the Department of Computer Science. He joined Lagos State University as an associate professor of Computer Science in 2013 and served as the pioneer departmental postgraduate coordinator. In 2015, he joined the University of Energy and Natural Resources as an associate professor and head of the Department of Computer Science and Informatics. He is currently a full professor at the Faculty of Computing, Engineering and Mathematical Sciences, Catholic University of Ghana, Sunyani. His research interest covers artificial intelligence, data mining, neural networks, and fuzzy logic. He can be contacted at email: adebayo.adekoya@cug.edu.gh.

Mighty Abra Ayidzoe holds a Ph.D. and a Master of Engineering in Software Engineering, which she received from the University of Electronic Science and Technology of China. She has a B.Sc. Computing with accounting degree from the University for Development Studies. Her current research interest covers deep learning, artificial intelligence, cybersecurity, internet of things, image processing, computer vision, artificial intelligence of things, and pattern recognition. She is a lecturer in the Department of Computer Science and Informatics at the University of Energy and Natural Resources, Sunyani, Ghana. She is currently with the Jamk University of Applied Sciences pursuing a second master's degree in Cybersecurity. She can be contacted at email: mighty.ayidzoe@uenr.edu.gh or mighty.ayidzoe@uds.edu.gh.