Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

An Empirical Evaluation and Comparative Study of Metaheuristic-Optimized Deep Learning for Four-Class Retinal Disease Classification

Audrey Huong[1], Wan Mahani Hafizah Wan Mahmud[2], Ser Lee Loh[3], Kim Gaik Tay[4], Xavier Ngu[5]

[1,2]Department of Electronic Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia.
[3]Centre for Robotics and Industrial Automation, Fakulti Teknologi dan Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia.
[4]Department of Computer Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia.
[5]RF EMC Centre Malaysia Sdn. Bhd., Universiti Tun Hussein Onn Malaysia, Johor, Malaysia.

E-mail: [1]audrey@uthm.edu.my, [2]wanmahani@uthm.edu.my, [3]slloh@utem.edu.my, [4]tay@uthm.edu.my, [5]xavier@uthm.edu.my

Orcid ID: [1]0000-0002-4505-5860, [2]0000-0001-9940-9292, [3]0000-0003-3870-2561, [4]0000-0002-7280-1363, [5]0000-0001-8718-9931

Abstract

Traditional retinal disease diagnosis using colour fundus imaging is subject to inter- and intra-observer variability and subjective bias. Meanwhile, manual or heuristic hyperparameter selection for neural network training is vulnerable to suboptimal convergence. There are also common challenges, including the lack of standardized evaluation protocols, non-uniform experimental designs, and class imbalance, which further hinder the development of reliable classifiers. This research addresses these limitations by employing metaheuristic algorithms to optimize the training of AlexNet, GoogleNet, and EfficientNet-B0. Particle swarm optimization (PSO), grey wolf optimization (GWO), and wild horse optimization (WHO) are evaluated for their effectiveness in identifying optimal training hyperparameter configurations for a four-class fundus image classification task. Statistical analysis revealed no significant association between the choice of optimization algorithm and the prediction performance (ρ = 0.341). The difference is more pronounced when testing the relationship between the network architecture and the prediction outcomes (ρ = 0.015). Among the evaluated optimizer-architecture pairs, GWO-EfficientNet demonstrates superior performance, achieving an accuracy of 96.52%, precision of 96.50%, recall of 96.47%, specificity of 98.84%, *F1* score of 96.48%, and a Matthew's correlation coefficient (*MCC*) of 95.33%, outperforming other optimizer-architecture combinations. Repeated experiments show strong performance consistency across the best-performing pairs, with a standard deviation of below 2%. The high accuracy and low inference time highlight the potential of the proposed approach for real-time ophthalmology applications, supporting improved clinical decision-making and more efficient eye care delivery.

Keywords: Deep Learning, EfficientNet, Fundus, Hyperparameter, Optimization.

DOI: https://doi.org/10.36548/jiip.2026.2.012
Received: 10.03.2026, received in revised form: 13.04.2026, accepted: 02.05.2026, published: 12.05.2026

1. Introduction

The eyes are essential organs important for visual perception. They are made of vitreous humor and neuroepithelial cells (i.e., photoreceptors), which contain numerous blood vessels necessary for cell sustainability [1]. Light entering the eye is focused onto the light-sensitive retina, where it is converted into electrical activity in photoreceptors. The signals are then transmitted via the optic nerves (OD) to the brain for visual perception. Age and health status play important roles in determining the functional and physical health of the eyes [2], and disorders of the eyes can stem from various factors, ranging from degradation of the vitreous media to inadequate nutritional and oxygen supplies required by the cells [3]. Some common eye diseases, such as diabetic retinopathy (DR), age-related macular degeneration, glaucoma, and vein and arterial occlusion, are the result of underlying illnesses in patients, including diabetes and inflammatory disorders [2].

Direct ophthalmoscope and fundus photography are the current methods used to examine and evaluate the internal structures of the eyes. These systems comprise a high-power light source, a magnification system, and an imager or camera, which enable the examination of microvascular patterns, haemorrhage, OD, retina, and macula in the cavity through real-time video streaming or recorded photographs [4]. This manual investigative process is highly subjective and poorly reproducible, as it depends on the medical experts' skills and experience, which may yield different results among and between observers. Recent advancements in the field include the use of optical coherence technology (OCT) [5] or acoustic-based approaches, such as Doppler and photoacoustic methods [6], for improved visualization. Iovino et al. [5] demonstrated the use of coherence interferometry to construct the topographical image of scanned tissues; however, the technique suffers from limited penetration depth and field of view. The acoustic method used in [6] can effectively visualize the deep structures of the eye, but it requires direct contact between the transducer and the eye; thus, the procedure can be uncomfortable for patients.

Following the growing trend in Artificial Intelligence (AI), many researchers in this domain have explored various AI approaches to perform instance classification or disease prediction from input images. This technology is widely used in the medical field to assist medical professionals in their diagnoses and increase their confidence in decision-making. Metaheuristic algorithms are AI search methods that have garnered significant attention in the last decade for their applications in global optimization research, owing to their robustness, efficiency, and enhanced model convergence [7]. Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Bayesian Optimization (BO) are some of the traditional metaheuristic optimization algorithms that randomly search for the best solution to a problem within the decision space, replacing the standard grid search, i.e., a brute-force exhaustive strategy that evaluates multiple parameter combinations to find the optimal solution. Several studies have reported the successful implementation of BO-enhanced CNN models, e.g., InceptionV4 and ResNet50 [8], for finding optimal training hyperparameter values.

Meanwhile, nature-inspired algorithms, such as the PSO, which mimics the social behavior of fish schooling and bird flocking, are subsets of metaheuristics computationally modeled after animals in their quest to survive and thrive in competitive environments. Ghosh et al. [9] proposed using the PSO to enhance the luminance and contrast of retinal fundus images, thereby increasing the visibility of retinal blood vessels and OD regions. Another effort using the same strategy in this domain investigated changes in MobileNet's performance by adjusting the learning rate and the number of epochs [10]. Their results show strong

performance in grading DR severity, with precision, recall, and F1-score ranging from 96% to 98%. Koishiyeya et al. [11] compared the performance of PSO, Artificial Bee Colony (ABC), and Binary Cuckoo Search (BCS) in optimizing feature dimensionality before classification is made using Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multi-layer Perceptron (MLP). The chronological tunicate swarm algorithm is another swarm-based algorithm introduced to optimize a deep Stacked Autoencoder (SAE) for DR severity classification [12]. In a separate study by Geetha et al. [13], the performance of glaucoma detection using an Aquila-optimized convolutional neural network (CNN) was investigated. The researchers observed an increase in disease localization detection performance of 3-22% using the employed swarm-based optimization algorithm compared to the attention-based CNN, the concatenated image features method, U-Net, and the conditional configuration-generative adversarial network (cGAN). Grey wolf optimization (GWO) is a new metaheuristic algorithm applied by [14] to optimize the segmentation threshold for localizing OD before the output is used for glaucoma classification using DenseNet201, VGG19, InceptionV3, and ResNet-50V2 in the subsequent stage. The research used the DRISHTI-GS dataset in their demonstration and reported performance metric scores ranging from 94 to 99% using hold-out validation and cross-validation. Instead of using GWO for the segmentation task, this optimization algorithm has also been used to identify the hyperparameter configurations, i.e., number of epochs, learner types, initial learning rate, and momentum, optimal for training DenseNet121, ResNet50, ResNet101V2, InceptionResNetV2, and Xception for detecting normal, glaucoma, and DR in [15]. The strategy was shown to enhance exploration efficiency and accelerate model convergence, resulting in overall high metric scores of 80-89%. Another study by [16] introduced the Red Spider Optimization (RSO) algorithm as a potential strategy to enhance the architectural design of an Active Gradient Deep Convolutional Neural Network (AG-DNN) by optimizing the model's learning parameters for classifying retinal features. The system was established based on extensive datasets from DRIVE, STARE, CHASE DB1, HRF, DRISHTI-GS, and RFMiD. The proposed system was also compared with Bayesian and Hummingbird optimization approaches and showed a slight improvement in accuracy of 1-2%. Wild Horse Optimization (WHO) and Elephant Herding Optimization (EHO) are some of the other approaches inspired by animals' complex social mechanisms used to solve similar medical challenges, such as glaucoma detection [17], [18] and retinal vessel segmentation [19], using the traditional deep learning models, including AlexNet, GoogleNet, VGG and ResNet, and their variants.

Despite advancements in optimization algorithm-enhanced computer-aided diagnosis systems, there remains a need to address inconsistencies in the problem formulation and the optimization approach used. Earlier studies in the field employed optimization methods for architecture exploration [12], [13], image quality enhancement [9], feature selection [11], and for identifying the best feasible training hyperparameter solution in the decision space [8], [15], Currently there is no clear consensus on the optimal network architecture or optimization framework for this task. These challenges are further complicated by variations in dataset-specific characteristics, inconsistent performance enhancement techniques, and the lack of standardized evaluation and experimental designs. The choice of optimization and deep learning methods was often made empirically, without justification or examination of alternative optimization strategies that may influence convergence stability or performance consistency. This paper aims to examine and compare the effectiveness of the PSO, WHO, and GWO algorithms across different models in enhancing model generalizability. Since existing experimental designs focus on simple, two-condition diagnostics or single-disease diagnostic systems that oversimplify real-world complexity, this study also investigates the performance

of optimization-enhanced systems on more complex retinal disease classification problems beyond those presented in [13], [16], [18]. The contributions of this study are manifold. First, it presents a systematic optimization-driven comparative analysis of deep learning architectures for fundus image classification. Second, this research introduces a more robust evaluation framework by formulating hyperparameter tuning as a four-dimensional nonconvex optimization problem, enabling an in-depth investigation of the proposed framework's performance. Third, the experimental findings provide new insights into the key factors influencing architectural effectiveness under different optimization strategies and identify the most suitable optimization–architecture pipeline for the automatic detection of common retinal diseases. All the following simulations and experimental analyses were conducted using MATLAB version 2025a.

2. Methods

Metaheuristic optimization is an advanced method for improving a model's learning efficiency by iteratively minimizing a loss function that depends on the quality of the chosen hyperparameters. There are, however, limited studies exploring differences in the performance of optimization methods and their application across various deep learning models. In the following subsections, this paper discusses the use of three nature-inspired metaheuristic algorithms, i.e., PSO, WHO, and GWO, that have been used to iteratively optimize the predictive performance of different CNN models for fundus image classification. These optimization methods employ different exploration-exploitation strategies; thus, their applicability and compatibility across networks could differ significantly. To ensure an unbiased and fair comparison of the optimization methods' robustness and performance, this paper used identical search configurations, experimental setups, and evaluation frameworks, as discussed in the following subsections.

2.1 Colour Fundus Photographs and Data Handling

This work used a publicly available dataset of fundus images for various eye diseases, accessible on the Kaggle website in [20], for the classification problem. This imbalanced dataset contains fundus images with four annotated retinal disease labels: normal, diabetic retinopathy (DR), cataract, and glaucoma, collected from different sources, i.e., the Indian Diabetic Retinopathy Image (IDRiD) dataset, High-Resolution Fundus (HRF) Image Database, Ocular Disease Recognition, Retinal Disease Classification dataset and Digital Retinal Images for Vessel Extraction (DRIVE) dataset. The images in these datasets were manually labelled by trained human readers. There are 1,039 images in the cataract class, 1,099 in the DR class, and 1,088 and 1,075 in the glaucoma and normal classes, respectively. Examples of these images are shown in Figure 1, which demonstrates distinct contrast and brightness variations even within the same class. Unlike earlier research that relied on noise filtering [14], [18] or image enhancement [9], [21] to improve the training results, no image manipulations or contrast enhancement have been carried out in this work to preserve the variability in image quality and characteristics, consistent with realistic conditions to ensure a more robust evaluation of the models' learning capability and predictive performance.

Figure 1. From Top Left to Bottom Right. Examples Of Fundus Images for Cataract, Diabetic Retinopathy (DR), Glaucoma, and Normal

The image data were collected from different institutions in various sizes and formats; thus, resizing is the only preprocessing step performed before training. Although downsizing images may degrade image quality and resolution, particularly the visibility of blood vessels, the macula, and the optic discs, this step is important to ensure consistent image input required for training the networks and for a direct, fair comparison of the employed models' performance. Hence, all images were scaled to a uniform size of 256 × 256 pixels before being split into training, validation, and test sets in a ratio of 0.8:0.05:0.15. A seed value of 1 was used to ensure a random yet repeatable selection of images within each dataset.

2.2 Classification Model

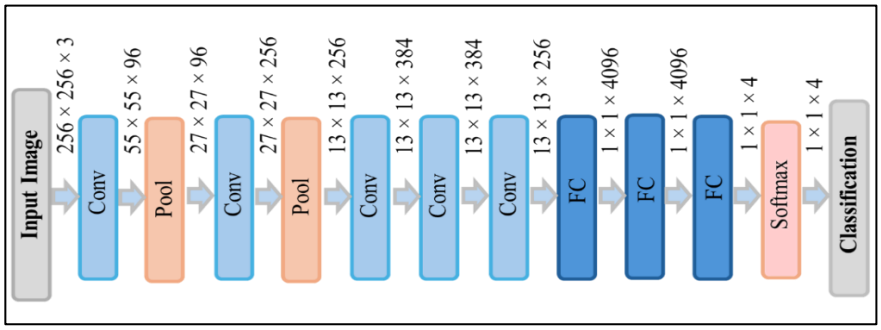
The fundus image classification was performed using CNN-based models. Our experiments used three important pretrained models for the task: AlexNet, GoogleNet, and Efficient-B0 models. Despite being among the first CNN models ever built, the eight-layer AlexNet, shown in Figure 2(a), remains popular today. It has been used as the standard CNN for transfer learning due to its efficiency and comparatively uniform, simple architecture. This model has not implemented depthwise separable convolutions or scaling strategies, making it a suitable early design baseline for straightforward comparisons with the competing GoogleNet and EfficientNet models.

GoogleNet, shown in Figure 2(b), is a 27-layer CNN consisting of nine inception modules (incept), which effectively increase the network's width and depth, and are important for capturing richer features in the input data while maintaining computational efficiency. This model also includes auxiliary classifiers to mitigate vanishing gradients. These characteristics make GoogleNet the first model to explicitly prioritize computational efficiency, serving as a conceptual predecessor to EfficientNet. The latter is the state-of-the-art model chosen in this paper due to its balanced trade-off between model size and performance [22]. EfficientNet-B0, shown in Figure 2(c), is the smallest model version in the EfficientNet family that uses the scaling method to uniformly scale all dimensions of network depth ($d\Phi$), width ($\omega\Phi$), and resolution ($r\Phi$). It employs Mobile Inverted Bottleneck Convolution (MBConv), consisting of a 1×1 expansion convolutional layer and a dropout layer (see the inset of the figure) for scaling the network's depth, width, and resolution using a constant compound coefficient, Φ. Increasing network dimensions enableS the network to capture more invariances and extract finer details, while higher-resolution feature maps enhance localization accuracy and thus can substantially improve network performance. The MBConv types used in Figure 2(c) varied with the convolutional filter size, k, i.e., 3 or 5. The final EfficientNet-B0 network consists of 16 MBConvs and is 237 layers deep. The input size of all networks in Figure 2 was set to 256 × 256 × 3 pixels for consistency in the comparison, and their Softmax layers were set to 1 × 1

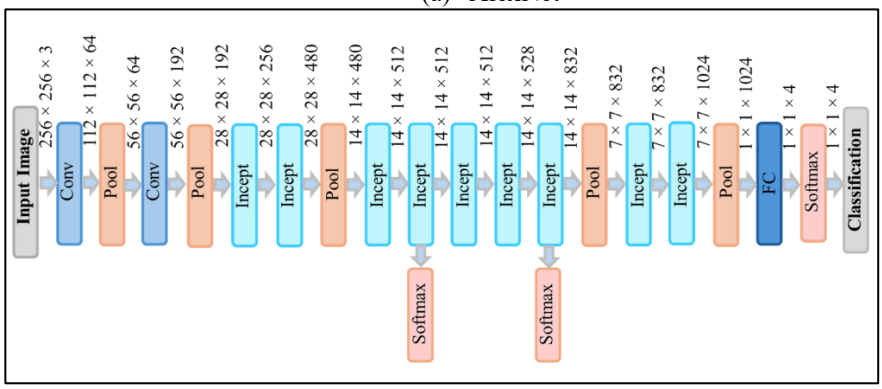
× 4, corresponding to the number of disease classes. Finally, the weights of the remaining layers across all networks were adjusted on the new dataset during training.

2.3 Nature-Inspired Optimization and Search Space

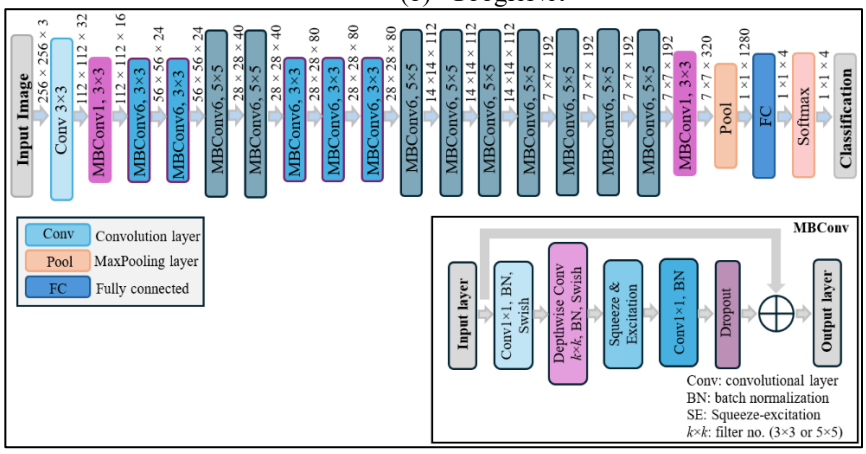
The models' generalizability can be improved by selecting suitable training hyperparameters to extract the most relevant features from the available data. The training parameters that significantly impact the model's performance include learner type (ζ), epoch number (N), mini-batch size (m), and initial learning rate (η), as recommended by [4], [9].



(a) AlexNet



(b) GoogleNet



(c) EfficientNet-B0

Figure 2. Network Architecture of (a) AlexNet, (b) GoogleNet, and (c) EfficientNet-B0 Modified for Fundus Image Classification

Instead of using the manual grid search method, this paper explores automatic optimization approaches for the problem. They include the conventional PSO, one of the most popular methods for finding global solutions, and the state-of-the-art WHO and GWO models,

which are well-known for solving complex, high-dimensional optimization problems, as illustrated in Figure 3. Each method determines the best combination of training hyperparameters by searching for the convergence point that produces the minimum cost function value in Eq. (1) within the four-dimensional search space defined in Table 1.

Table 1. Four-Dimensional Search Space Boundaries

Hyperparameter	Search space	
	Lower limit	Upper limit
Learner type, ζ	1→3 {*Adam*, *Sgdm*, *RMSProp*}	
Epoch number, N	50	350
Mini-batch size, m	8	300
Initial learning rate, η	$1e^{-5}$	$1e^{-1}$

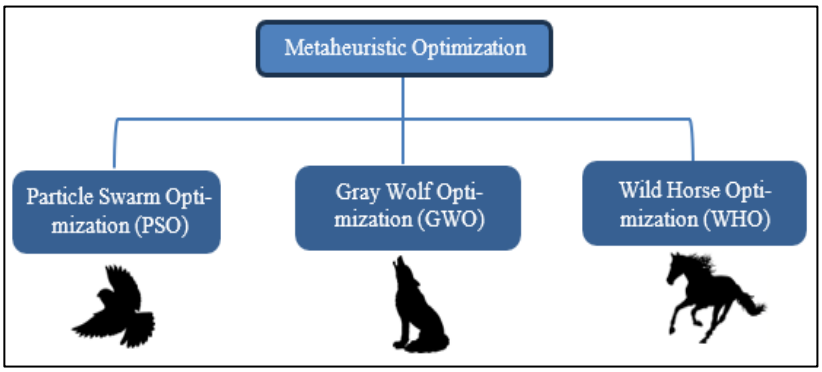


Figure 3. Animal Behavior-Inspired Optimization Methods for Optimizing Model Training Efficiency and Prediction Accuracy

For consistency in our investigation, this paper used a population count of 20 for each optimization approach (i.e., PSO, GWO, and WHO) in training the models shown in Figure 2. The experiments ran one approach after another on each model, on the same machine (NVIDIA Tesla K80 GPU, 256 GB RAM). The optimization process began by randomly initializing candidate solutions in the four-dimensional search spaces shown in Table 1, using a seed value of 1. The performance of each particle (i.e., candidate solution) was evaluated using the objective function, f, in Eq. (1).

$$f(T_{acc}, V_{acc}, t) = (1 - T_{acc}) \times 10^2 + (1 - V_{acc}) \times 10^3 + {}^{t}/_{1000} \quad (1)$$

T_{acc}, V_{acc}, and t represent training and validation accuracies and training time, respectively. The accuracy metrics (T_{acc}, V_{acc}) and the computing time (t) are collectively important in evaluating model convergence efficiency. In this paper, t is normalized by a factor of 1000 to prevent the time component from dominating the objective function, bringing its numerical magnitude to a scale comparable to that of the accuracy-related terms. Since V_{acc} is a reliable indicator of a model's ability to generalize to unseen data, while T_{acc} reflects data fitting, a higher penalty was applied to the validation set prediction error to address the issue of model convergence to suboptimal solutions. This iterative evaluation-search process can be illustrated in Figure 4. An initial set of randomly generated hyperparameter configuration is used for model training. The resulting performance measures (T_{acc}, V_{acc}, and T_s) from the training session guide the optimization at each iteration through minimization of the objective function in Eq. (1). The solution quality influences training efficiency and, consequently, the objective function value; this nonlinear dependency renders the relationship between the minimization problem and the solution implicit.

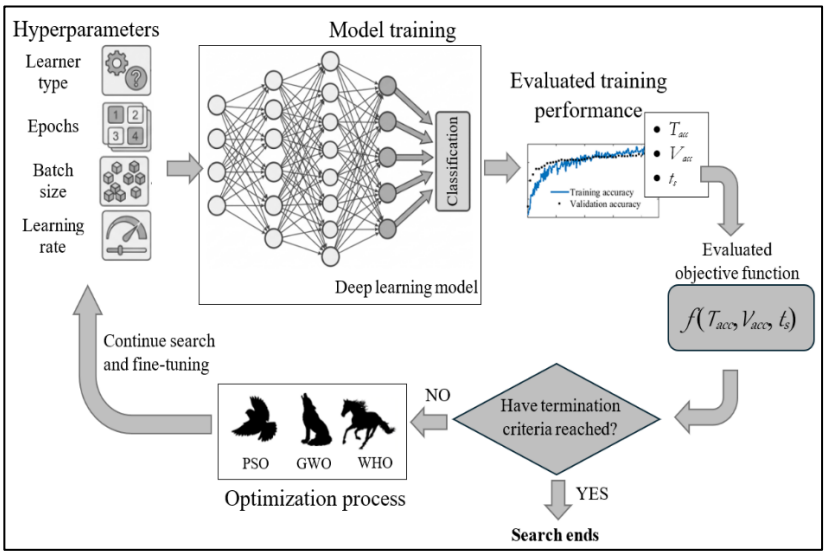


Figure 4. Nonlinear Relationship between the Optimization Process and Hyperparameter Solutions

2.3.1 Particle Swarm Optimization (PSO)

PSO is an efficient neighbourhood-based search strategy involving a swarm of search agents (i.e., particles or candidate solutions) that remember their own best position (pbest) and the swarm's best position (gbest). These particles explore the solution space by adjusting their position based on the quality of their solutions compared to those of their neighbors. After releasing a specified number of particles into a search area, PSO iteratively searches for an optimal solution. The position, x, and velocity, v, of each particle are recorded at each iteration, and the position is evaluated for its fitness value. The current best solution corresponds to the point (i.e., position) that produces the fittest solution to the objective function in Eq. (1) at the nth iteration. After the best solution is identified, particles from the entire population would update their velocity by moving towards the best solution point in the iteration that follows, i.e., (n+1). Equation (2) shows the updated velocity for particle i, using which its new position, xi, can be determined.

$$v_i(n+1) = w \cdot v_i(n) + c_1 r_1\big(pbest_i - x_i(n)\big) + c_2 r_2\big(gbest_i - x_i(n)\big) \quad (2)$$

$$x_i(n+1) = x_i(n) + v_i(n+1) \quad (3)$$

This paper used the default inertia weight range, w = [0.1, 1.1], and the term w.vi in Eq. (2) represents the particle's momentum that preserves the particle's movement vector from the previous position. The c1 and c2 are weighting factors for self and social adjustment, respectively, and have the same default value of 1.49, whereas r1 and r2 are random variables ranging from 0 to 1. These second and third terms in Eq. (2) represent randomization processes that affect the exploitation and exploration potential of the particles, i.e., around their own best past position, pbest, or the best position found of the swarm, gbest, respectively. The current best solution, xi, is updated to the new best solution in Eq. (3), determined using the velocity updated after each iteration.

2.3.2 Grey Wolf Optimization (GWO)

GWO is a structured leadership-hierarchical search method that works by identifying the three best solutions, i.e., alpha (α, the fittest solution evaluated using Eq. (1)), beta (β, the

second-best solution), and delta (δ, the third-best solution), from a group of wolves (i.e., solutions). The algorithm starts by randomly initializing the position of a population of grey wolves (i.e., candidate solutions) within the search space, before the three leaders (i.e., the three best solutions) are identified based on the evaluated fitness, which guides the remaining candidate solutions in searching for the prey (i.e., global optimum). The process begins by estimating the distance vector $\vec{D}$, which measures the distance between a wolf $\vec{X}(n)$ and the prey position vector $\overrightarrow{X_p}(n)$ in the current iteration, n, as shown in Eq. (4).

$$\vec{D} = \left|\vec{C} \cdot \overrightarrow{X_p}(n) - \vec{X}(n)\right| \tag{4}$$

$\vec{C}$ is a random coefficient introduced to promote movement diversity, given by:

$$\vec{C} = 2 \cdot \vec{r} \tag{5}$$

$\vec{r}$ is a random vector ∈ [0, 1]. The position vector of the search individuals is updated in the subsequent iteration, $\vec{X}(n+1)$, in Eq. (6) using the distance vector estimated in Eq. (4). The term $\vec{A} \cdot \vec{D}$ in Eq. (6) is the position adjustment vector that determines the wolves' movement sizes, where the coefficient $\vec{A}$ ranges between [-a, a] controls the exploration-exploitation behavior of the wolves. The value a decreases from 2 to 0 throughout the iterations, representing the transition from exploration to exploitation (convergence).

$$\vec{X}(n+1) = \overrightarrow{X_P}(n) - \vec{A} \cdot \vec{D} \tag{6}$$

The leadership hierarchy (i.e., the positions of α, β, and δ members) is adaptively updated at every iteration, enabling stable convergence toward the global optimum.

2.3.3 Wild Horse Optimization (WHO)

The WHO is a population-based approach that begins by randomly generating an initial population of horses. Similar to the other optimization algorithms, the position of each horse is evaluated for its fitness using the objective function in Eq. (1). The population is then divided into four groups, each led by a group leader, i.e., a stallion represented by the symbol Sn, which has the best fitness function in the group. The rest are foals that would move toward their stallion during the exploitation phase. The position of the i-th search individual, i.e., foal, Xi, that grazes around the stallion's position is updated to a new position at n+1 iteration using Eq. (7).

$$X_i(n+1) = V_R \cdot \left(S^n - X_i(n)\right) + S^n \tag{7}$$

Where VR denotes the movement coefficient that controls the direction and the step size the foal moves relative to its stallion, allowing oscillatory exploitation, defined as follows:

$$V_R = 2Z\cos(2\pi RZ) \tag{8}$$

R is a random number in the range [-4, 4], and Z is a dimension-dependent random vector in the range of [0, 1]. The new position of each group is evaluated, and the candidate solution that outperforms the leader becomes the new stallion. This algorithm also includes a crossover operation to improve global search capability, with the foal crossover rate (CR) to another group arbitrarily set to 0.13. At each iteration, a random number uniformly distributed in [0, 1] is generated and compared with the CR. If this value is less than the CR, the worst-

performing foals from two randomly selected stallions are averaged, producing a new candidate solution.

The location of each candidate solution in Sections 2.3.1- 2.3.3 is updated after each iteration. The above-described procedure is iterated five times and terminated early when validation accuracy fails to increase after ten training iterations or when the maximum epoch number is reached. This process is applied and repeated for all considered networks in Figure 2.

2.4 Evaluation and Performance Analysis

The classification performance of the optimized networks was evaluated on the unseen test dataset, and their quality was assessed using the common performance metrics shown in Eqs. (9)-(13). These metrics include accuracy (ACC), which measures the number of accurate predictions divided by all predictions; precision (PREC) and recall (REC) or sensitivity, which measure the quality and quantity of positive predictions, respectively; and specificity (SPEC), which evaluates the accuracy of negative predictions. This study also employed the F1 score, which combines precision and recall, to assess the system's quality and to detect differences between the classes. The MCC, which provides a robust statistical correlation measure for imbalanced datasets like ours, and the AUC, which reflects the overall discriminative ability of the models, are also used to evaluate the models' generalizability. These metrics are given in Eqs. (14) and (15), respectively.

$$ACC = \frac{(TP+TN)}{(FN+FP+TP+TN)} \tag{9}$$

$$PREC = \frac{TP}{(TP+FP)} \tag{10}$$

$$REC = \frac{TP}{(TP+FN)} \tag{11}$$

$$SPEC = \frac{TN}{(TN+FP)} \tag{12}$$

$$F1 = \frac{2 \cdot PREC \cdot SENS}{PREC+SENS} \tag{13}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{14}$$

$$AUC = \int_0^1 SENS \, d\left(\frac{FP}{FP+TN}\right) \tag{15}$$

TP denotes the true positive rate, i.e., the number of instances correctly predicted as positive out of all positives, and TN denotes the correct prediction of negative cases out of all true negatives. FP is the number of instances falsely predicted as positive out of all true negatives, whereas FN is the number of instances misclassified as negative out of all true positives.

3. Results and Analysis

The iterative process of optimizing the training of AlexNet, GoogleNet, and EfficientNet-B0, shown in Figure 2, is assumed to return the best solution identified from the

search space defined in Table 1. Shown in Figure 5 is the testing confusion matrix for the models trained with the best solution (i.e., the training hyperparameter configuration) determined by the PSO, GWO, and WHO algorithms through the iterative search described in the earlier section. Class labels I-IV represent cataract, DR, glaucoma, and normal classes. The performance measures calculated based on these confusion matrices are tabulated in Table 2. Also shown in the table is the total elapsed time, Ts, in seconds (s) recorded for the search operation. The overall AUC values for the best-performing WHO-AlexNet and both PSO-GoogleNet and GWO-EfficientNetB0 models are 0.996 and 1, respectively. The best testing results for each optimized network, referred to in the remainder of the paper as the optimizer-architecture pair, are highlighted in bold, and their respective training progress plots (i.e., WHO optimized AlexNet, PSO-GoogleNet, and GWO-EfficientNetB0) are illustrated in Figure 6 for the completeness of our presentation. The present study did not directly compare the obtained results with the literature due to the differences in the evaluation processes and datasets used in the experiments, wherein earlier studies in the field by [13], [18] were limited to simpler tasks of binary classification, i.e., classification of normal and abnormal fundus images, or ternary applications [15], or only targeted a single disease [8], [12].

A one-way analysis of variance (ANOVA) was performed using IBM SPSS Statistics for Windows, Version 22.0, to compare the performance scores presented in Table 2 across the three optimization methods. The difference in their means was found to be statistically insignificant, with a ρ value of 0.341, at the 95% confidence level. Additionally, data analysis conducted using the same statistical method to evaluate differences in prediction results among deep learning models revealed a statistically significant result with a ρ value of 0.015. The best hyperparameter set, {ζ, N, m, η}, identified for AlexNet using WHO is {Adam, 64, 170, 1e-5}. The best-performing PSO-optimized GoogleNet and GWO-optimized EfficientNet-B0, i.e., PSO-GoogleNet and GWO-EfficientNet-B0, chosen based on the results in Table 2, were trained using hyperparameter sets of {Sgdm, 280, 274, 0.0018} and {Adam, 226, 64, 1e-3}, based on the findings in Table 2.

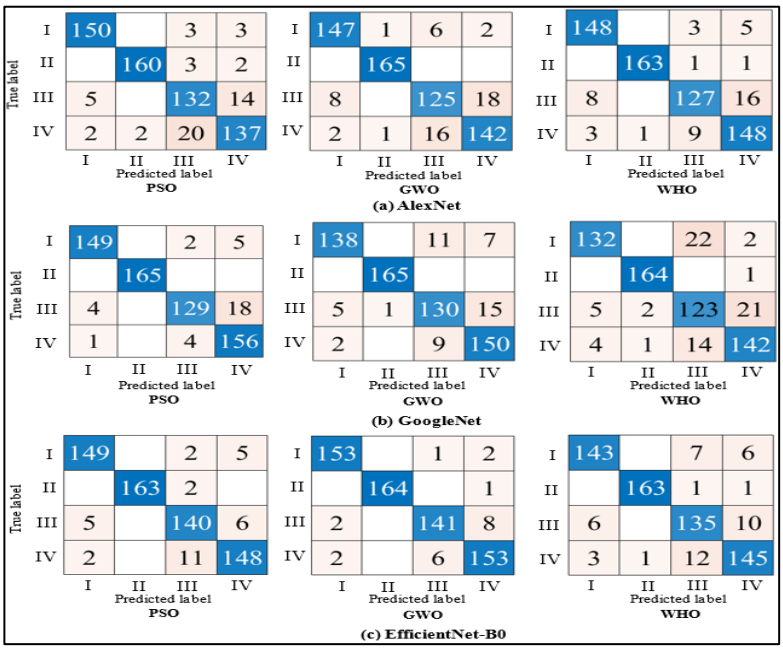


Figure 5. Testing Confusion Matrix of (a) AlexNet, (b) GoogleNet, and (c) EfficientNet-B0 Optimized Using Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), and Wild Horse Optimization (WHO) for Retinal Disease Classification (Label I: Cataract, II: Diabetic Retinopathy, III: Glaucoma, IV: Normal)

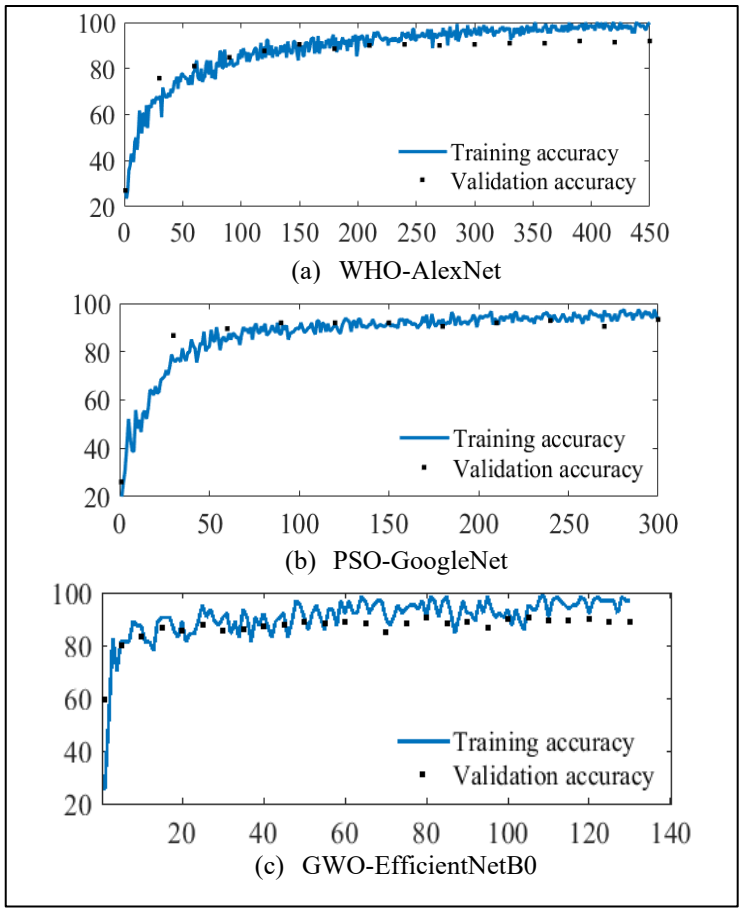


Figure 6. Training Progress Plot of the Best-Performing (a) WHO-AlexNet, (b) PSO-GoogleNet, and (c) GWO-EfficientNetB0. The Graph Illustrates the Training (Blue Line) and Validation Accuracies (Black Square Dots) at Each Iteration

Table 2. Evaluated Performance Measures of Models Optimized Using Nature-Inspired PSO, GWO, and WHO Methods

Model	Optimization algorithm	Performance metric (in %)[a]						
		ACC	PREC	REC	SPEC	F1	MCC	T_s(s)
AlexNet	PSO	91.47	91.42	91.41	97.17	91.39	88.58	$2.5e^4$
	GWO	91.47	91.28	91.30	97.17	91.29	88.47	$1.9e^4$
	WHO	92.58	92.56	92.42	97.53	92.44	90.02	$2.8e^4$
GoogleNet	PSO	94.63	94.86	94.46	98.21	94.53	92.86	$3.5e^4$
	GWO	92.10	92.11	91.93	97.38	91.97	89.39	$1.7e^4$
	WHO	91.31	91.18	91.12	97.11	91.06	87.76	$4.8e^4$
EfficientNet-B0	PSO	94.79	94.73	94.74	98.27	94.73	93.00	$1.3e^5$
	GWO	96.52	96.50	96.47	98.84	96.48	95.33	$1.3e^5$
	WHO	92.58	92.52	92.48	97.54	92.49	90.03	$1.1e^5$

[a]*ACC*: accuracy; *PREC*: precision, *REC*: recall, *SPEC*: specificity, *MCC*: Matthew's Correlation Coefficient, T_s: Elapsed time

This paper also conducts class-wise analysis to explore the model's detection performance across different disease classes, providing a better understanding of the model's effectiveness and limitations. The precision, recall, and F1 scores of the best-performing models in Table 2 are analyzed to investigate their per-class classification performance, as tabulated in Table 3. Their per-class Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) and the optimal operating point, denoting the threshold at which the model achieves the best balance between the recall (true positive rate, TPR) and false positive rate (FPR), are shown in Figure 7. Also shown in the figure are their respective precision-recall

(PR) curves, which evaluate the trade-off between precision (correct retinal class detection) and recall (ability to detect all retinal classes). The plots illustrate higher AUCs for both ROC and PR curves, especially for GWO-EfficientNetB0, with diabetic retinopathy detection achieving near-perfect to excellent performance (0.999-1) across all optimizer-architecture pairs.

Table 3. Comparison of Class-Wise Precision (PREC), recall (REC), and F1 Score of the Best Performing Optimizer-Architecture Pairs Across the Retinal Disease Categories

Optimizer-architecture pair	Metrics	Retinal disease categories			
		Cataract	**Diabetic retinopathy**	**Glaucoma**	**Normal**
WHO-AlexNet	*PREC*	93.08	99.39	90.71	87.06
	REC	94.87	98.79	84.11	91.93
	F1	93.97	99.09	87.30	89.43
PSO-GoogleNet	*PREC*	96.75	100.00	95.56	87.15
	REC	95.51	100.00	85.43	96.90
	F1	96.12	100.00	90.20	91.80
GWO-EfficientNet	*PREC*	97.45	100.00	95.27	93.29
	REC	98.08	99.39	93.38	95.03
	F1	97.76	99.70	94.32	94.15

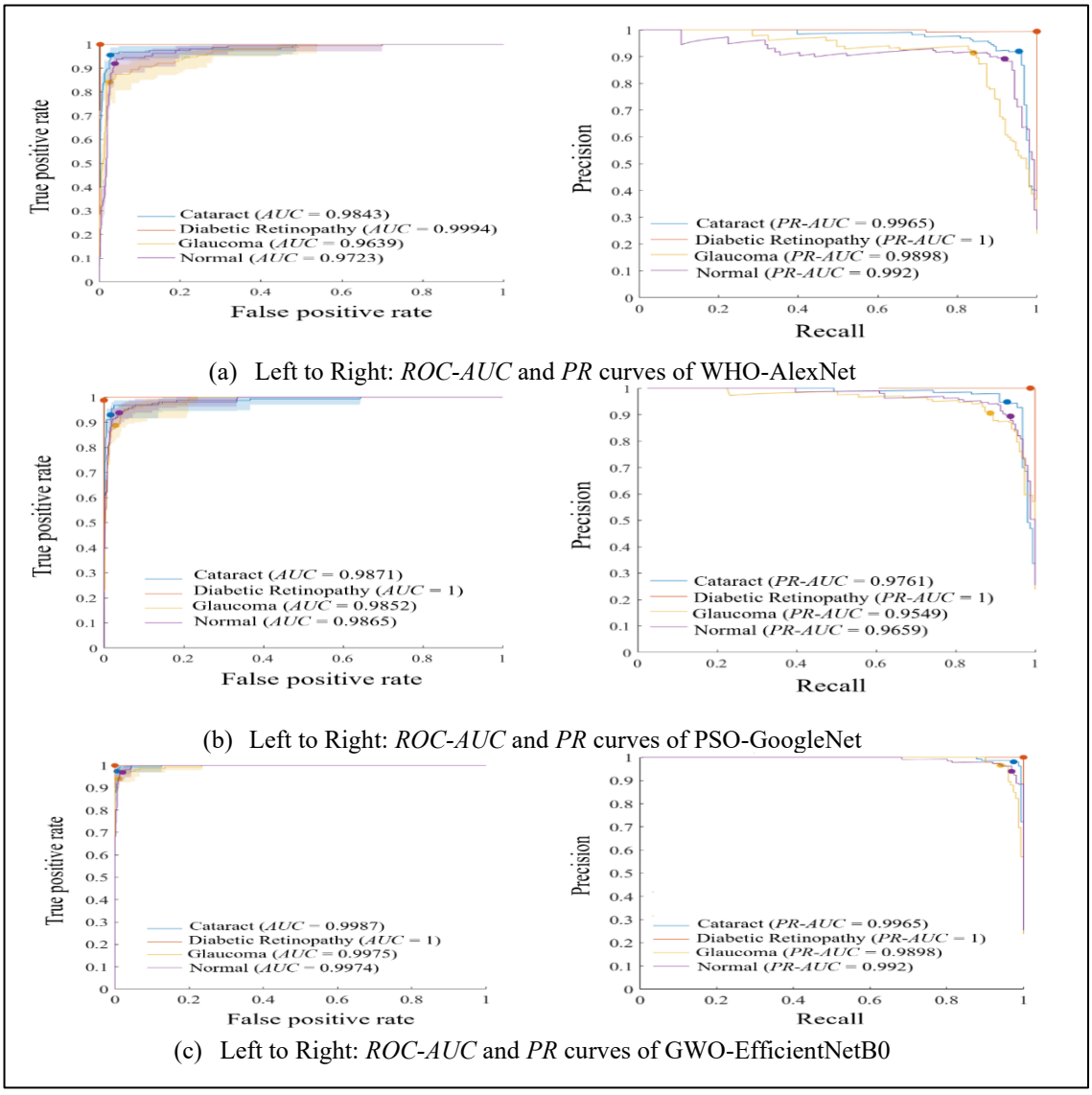


Figure 7. Class-Wise Receiver Operating Characteristic-Area Under Curve (ROC-AUC) and Precision-Recall (PR) Plots of Different Optimal Optimizer–Network Pairings

To assess performance consistency and reproducibility, each experiment was repeated twice, yielding a total of three runs for the best optimizer-architecture pairs identified in Table 2. The experiments used identical randomly selected initializations explained in Section 2.3. The repeated searches produced different optimal solutions for each model: {Sgdm, 305, 280, 0.0058} and {RMSProp, 195, 294, 1e-5} for WHO-AlexNet; {Adam, 178, 193, 1e-4} and {RMSProp, 131, 251, 1e-5} for PSO-GoogleNet; and {Sgdm, 135, 52, 0.0135} and {Sgdm, 64, 205, 0.0247} for GWO-EfficientNet. The results from the three runs are averaged, and their means and standard deviations are presented in Table 4.

Table 4. Mean and Standard Deviation of Performance Metrics from Three Independent Optimization Runs for the Selected Optimizer-Architecture Pairs

Optimizer-architecture pair	Performance metric (in %)					
	ACC	PREC	REC	SPEC	F1	MCC
WHO-AlexNet	92.2±0.5	92.1±0.6	92.1±0.5	97.4±0.2	92.1±0.5	89.5±0.7
PSO-GoogleNet	93.9±1.1	93.9±1.4	93.7±1.0	98.0±0.4	93.7±1.1	91.8±1.6
GWO-EfficientNet	96.5±0.7	96.4±0.7	96.4±0.7	98.8±0.2	96.4±0.7	95.3±1.0

4. Discussion

The traditional method of identifying the appropriate training hyperparameters to enhance model learning efficiency using grid search is exhaustive and inefficient, and it does not guarantee finding an optimal solution. Although nature-inspired optimization methods have been used for numerous retinal image decision-making tasks, such as image processing [9], [21] and OD and blood vessel segmentation [12], [13], [14], their use for the multi-class fundus image classification problem using different deep learning models has yet to be explicitly studied and compared. Instead of evaluating the efficiency of an optimization algorithm in combination with different architectures, as demonstrated by [15], [18], this paper explores and compares the performance of different optimization algorithms with varying existing models for fundus image classification. While PSO is widely used as a swarm-based method, both GWO and WHO are popular for solving complex, high-dimensional problems [23], such as ours, which involves four unknowns. The results of applying these optimization algorithms to various deep learning architectures demonstrated sufficient model convergence, independent of class imbalance, producing near-perfect performance across all evaluated metrics, with values ranging from 90% to 100%, as shown in Table 2. Their effectiveness is further supported by calculated MCC scores of 88-95% and comparable, near-perfect discriminative performance, with an overall AUC of 0.996-1. These experimental results underscore the superior performance of the optimization-empowered deep learning model, with notable improvements in evaluation metrics, outperforming earlier studies [8], [15], [22] that used traditional machine learning methods for binary and ternary classification tasks. This study found no competitive advantage among optimization techniques, $\rho = 0.341$; each method performs better in different models, and a different optimal hyperparameter set is identified for each model.

Although their performance is similar, based on the metric scores in Table 2, the WHO algorithm performed best at optimizing AlexNet, whereas PSO optimized GoogleNet and GWO optimized the training of EfficientNet-B0. Further comparisons of the prediction capabilities of these optimizer-architecture pairs revealed that GWO-EfficientNet achieved superior performance, while WHO-AlexNet performed the least efficiently, with a marginal 4% decline in classification accuracy. These results were confirmed by statistical significance

($\rho < 0.05$) between the models, implying that architecture effectiveness also depends on the adopted optimization approach. The training progress of the models shown in Figure 6 reveals distinct learning curves. AlexNet has fewer layers and a simpler structure, leading to smoother, stable training and faster convergence, as shown in Table 2. Therefore, the WHO, which has strong stochastic and oscillatory exploration capabilities to avoid local minima, is shown to be most appropriate for a simple network like AlexNet. On the contrary, complex networks like GoogleNet and EfficientNet that adopted multi-path structures and compound scaling methods, respectively, exhibit strong regularization effects that promote efficient learning. The GWO-EfficientNet showed rapid learning and adaptation, reaching high accuracy in fewer iterations (i.e., 20 iterations), indicating the model's efficiency in capturing intricate and subtle patterns essential for accurate prediction. However, this comes at the expense of longer convergence times, approximately five times longer than those of AlexNet. The computational complexity of this network also explains why higher initial learning rates of 1e-2-1e-3 were used. Nonetheless, a small variance in the gradient updates of this model can lead to fluctuations and introduce randomness into the training process, as shown in Figure 6(c). Despite these differences, the validation accuracies of all models closely track the training accuracy, suggesting good model generalization. The classification performance of GWO-optimized EfficientNet stands out with an impressive accuracy of 96.52% and high specificity (98.44%), making it the best-performing model for fundus image classification. The precision and recall scores of 96.50% and 96.47%, and the F1 score of 96.48%, also suggest effective detection of true positives. Notably, EfficientNet consistently outperforms competing models across different optimization algorithms. These results suggest that convergence efficiency is primarily governed by network architecture, and the considered optimization algorithms have comparable search performance, with little influence on predictions in efficient models such as EfficientNet-B0. Performance may be further enhanced by adopting deeper EfficientNet variants, i.e., B1-B7, which employ growing compound scaling to capture more complex feature representations. However, optimization dynamics may change as model depth increases, and different hyperparameter configurations may be identified even within the same tuning space.

This variability in search outcomes is evident in Table 4, obtained using the optimization framework described in Section 2.3. These algorithms incorporate stochasticity to enhance search efficiency, leading to different optimal solutions at the end of the iterations. For example, despite identical initialization points, different optimal hyperparameter sets, {ζ, N, m, η} of {Adam, 226, 64, 1e-3}, {Sgdm, 135, 52, 0.0135}, and {Sgdm, 64, 205, 0.0247} were obtained for GWO-EfficientNet across three runs. Randomness was introduced in PSO via stochastic values assigned to cognitive and social components (r1 and r2) in Eq. (2) to mitigate local minima and premature convergence. In GWO, the random coefficient vector $\vec{C}$ and control parameter $\vec{A}$ in Eqs. (5) and (6) promote probabilistic variation in encircling behavior, balancing exploration and exploitation. Similarly, stochastic horse grazing behavior in WHO is modelled in Eq. (8) to enhance search performance. The combined effects of stochastic optimization, inherent training randomness, the indirect nature of the minimization problem, as shown in Figure 4, and the high dimensionality of the search space contribute to variability in the resulting hyperparameter configurations across runs. Nevertheless, the strong performance consistency across three evaluation runs in Table 4 indicates the robustness of the search process in a multi-modal solution landscape where multiple near-optimal solutions may exist. High mean metric values (~91-98%) and low standard deviations (<2%), together with consistent performance trends, further demonstrate reliable convergence to high-performing regions of the search space. Moreover, a previous study by the same authors [24] reported

comparable performance between these optimization strategies and a hierarchical global-local optimization framework, but with significantly lower computational complexity.

The detailed analysis of their class-wise detection performance in Table 3 showed that GWO-EfficientNet achieved higher overall class separability, with the evaluated metrics ranging from 93 to 100%. Their ROC plot further revealed that the model consistently ranks true positives (TPR) higher than true negatives (FPR), as shown in Figure 7. The PR plot showed that the class-wise curves were closest to the upper-right corner of the PR space, with precision consistently exceeding 95% even at higher recall levels, demonstrating its superiority in positive-class detection. Further insights into the diagonal cells of the confusion matrices in Figure 5 and the ROC and PR curves in Figure 7 reveal the highest overall prediction measures for DR image classification, with all three CNN models accurately predicting most images with this class label. A near-perfect detection performance of 99-100% was determined for this class in Table 3. This is closely followed by the cataract, which showed AUCs of 0.97-0.99 in Figure 7 and performance metrics of 93-98% in Table 3. The glaucoma and normal classes exhibit higher FP and FN rates, compromising detection performance and resulting in lower precision and recall scores of 84-87%, especially in WHO-AlexNet. One possible explanation is that the DR fundus photograph often presents cotton-wool spots and macular oedema, which contribute to macular thickening, making it distinctive from other eye diseases. These are likely the features that all models learned equally well. Meanwhile, the cataract fundus photograph shows fewer details of the optic structure, depending on disease severity, which is a prominent feature for accurate classification. In contrast, the classification of glaucoma and normal images, especially with AlexNet and GoogleNet, is considerably inferior. There exists a high misclassification rate between these two classes. Clinical diagnosis of glaucoma is based on the size of the optic cup and disc, their ratio, and neuroretinal rim thickening [25], which can be hard to detect using images of limited resolution, as shown in Figure 1. Hence, this may be a contributory factor to the high misclassification rate between DR and normal fundus. The magnitude of this problem is less pronounced in the EfficientNet model, which achieved higher predictive accuracy for these retinal classes than its competitors.

Despite the long convergence time (Ts) reported in Table 2, the average inference time per image is measured in fractions of a second. The final trained model size of the GWO-EfficientNet is 14.5 MB. The short inference time and the small model memory footprint enable real-time prediction of fundus images captured by a portable fundus camera. Such a system can be used and extended to optometry to assist ophthalmologists in their diagnosis, increase confidence in decision-making, and optimize the delivery of eye care. Future efforts also include adopting larger, richer datasets of these classes and other retinal disease classes and conditions to enable a more comprehensive examination of different eye diseases. Since the results show that different optimization strategies perform more effectively in different models, there is considerable potential to extend this research by combining or hybridizing optimizers to achieve a more robust and efficient exploration of the search space.

5. Conclusion

This study used and compared particle swarm optimization, grey wolf optimization, and wild horse optimization to enhance the learning capabilities of several key deep learning networks, namely AlexNet, GoogleNet, and EfficientNet-B0, by searching for the optimal hyperparameter settings. The results demonstrated that the optimization methods can also be a viable solution for effectively managing imbalanced class problems. Although the experiments

found no significant association between the classification outcome means and the optimization algorithms used, a strong relationship was observed between the network architecture and the prediction outcome. The findings suggest the possibility of multiple optimal solutions in the decision space, leading to different optimal hyperparameter configurations across runs. Nonetheless, the comparable predictive capability of the optimized models confirms the robustness and efficiency of the search process. The GWO-EfficientNet is shown to outperform the others, achieving exceptional overall performance and a strong accuracy of 96.52%, suggesting its suitability for fundus image classification problems, which this research aims to address. This automated technique represents a new direction for enhancing the performance of the classification model and promoting the practical implementation of AI in diagnosis. Future work includes adding additional retinal disease datasets and possibly hybridizing optimization algorithms to promote rapid feature learning and enable robust performance across varying datasets.

Acknowledgments

This research was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2024/TK07/UTHM/02/5).

Declaration Statement

All authors declare that they have no conflict of interest related to this work.

References

[1] Jaffet, Jilu, Tejaswini Pingali, Arun Kumar Raut, Sonali Mohapatra, and Vivek Singh. "Eye: Anatomy, Physiology, and Disease." In Complex Ophthalmic Dosage Forms: Advances in Biomedical Applications and Future Perspectives, Singapore: Springer Nature Singapore, 2025, 45-69.

[2] Pieńczykowska, Kamila, Anna Bryl, and Małgorzata Mrugacz. "Link between Metabolic Syndrome, Inflammation, and Eye Diseases." International Journal of Molecular Sciences 26, no. 5 (2025): 2174.

[3] Pardeshi, Sagar R., Mahesh P. More, Abhijeet D. Kulkarni, Chandrakantsing V. Pardeshi, Pritam B. Patil, Ankit S. Patil, Prabhanjan S. Giram et al. "Current Perspectives in Nanomedicine Delivery for Targeted Ocular Therapeutics." Bulletin of Materials Science 46, no. 1 (2023): 35.

[4] Robles, Rafael, Nikhil Patel, Emily Neag, Ajay Mittal, Zahra Markatia, Kambiz Ameli, and Benjamin Lin. "A Systematic Review of Digital Ophthalmoscopes in Medicine." Clinical Ophthalmology (2023): 2957-2965.

[5] Iovino, Claudio, Clemente Maria Iodice, Danila Pisani, Luciana Damiano, Valentina Di Iorio, Francesco Testa, and Francesca Simonelli. "Clinical Applications of Optical Coherence Tomography Angiography in Inherited Retinal Diseases: An Up-To-Date Review of the Literature." Journal of Clinical Medicine 12, no. 9 (2023): 3170.

[6] Adenigba, Peter T., Ademola J. Adekanmi, and Olufunmilola A. Ogun. "Central Retinal and Ophthalmic Artery Doppler Velocimetry among Hypertensives and Normotensive Adults at a Nigerian Tertiary Health Facility." Nigerian Medical Journal 63 (2022):385–393.

[7] Dragoi, Elena Niculina, and Vlad Dafinescu. "Review of Metaheuristics Inspired from the Animal Kingdom." Mathematics 9, no. 18 (2021): 2335.

[8] Patil, Mahesh, Satyadhyan Chickerur, Vijayalakshmi Bakale, Shantala Giraddi, Vivekanand Roodagi, and Yashaswini Kulkarni. "Deep Hyperparameter Transfer Learning for Diabetic Retinopathy Classification." Turkish Journal of Electrical Engineering and Computer Sciences 29, no. 8 (2021): 2824-2839.

[9] Ghosh, Swarup Kr, Biswajit Biswas, and Anupam Ghosh. "A Novel Approach of Retinal Image Enhancement Using PSO System and Measure oof Fuzziness." Procedia Computer Science 167 (2020): 1300-1311.

[10] Raza, Asif, Shahrulniza Musa, Ahmad Shahrafidz Khalid, Muhammad Mansoor Alam, Mazliham Mohd Su'ud, and Fouzia Noor. "Multiclass Diabetic Retinopathy: Hybrid Metaheuristic Particle Swarm Optimization and Classification for Severity Grading and Feature Extraction." Engineering, Technology & Applied Science Research 15, no. 6 (2025): 30317-30323.

[11] Koishiyeva, Dina, Kuanysh Alipbayev, Jeong Won Kang, Adil Mukhamedgali, and Assel Mukasheva. "Optimisation of Glaucoma Detection in Fundus Imaging Using Particle Swarm Optimization, Artificial Bee Colony, and Binary Cuckoo Search." In 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), IEEE, 2025, 1-8.

[12] Dayana, A. Mary, and WR Sam Emmanuel. "An Enhanced Swarm Optimization-Based Deep Neural Network for Diabetic Retinopathy Classification in Fundus Images." Multimedia Tools and Applications 81, no. 15 (2022): 20611-20642.

[13] Geetha, A., M. Carmel Sobia, D. Santhi, and A. Ahilan. "DEEP GD: Deep Learning Based Snapshot Ensemble CNN with EfficientNet for Glaucoma Detection." Biomedical Signal Processing and Control 100 (2025): 106989.

[14] Almeshrky, Hamida, and Abdulkadir Karacı. "Optic Disc Segmentation in Human Retina Images Using a Meta Heuristic Optimization Method and Disease Diagnosis with Deep Learning." Applied Sciences 14, no. 12 (2024): 5103.

[15] Gül, Muhammed Furkan, Özlem Polat, and Halit Bakir. "TL-GWO: Fine-Tuned Transfer Learning with Grey Wolf Optimizer for Accurate Fundus Image-Based Eye Disease Classification." Experimental Eye Research (2025): 110598.

[16] Subramaniam, Krishnakumar, and Archana Naganathan. "Enhancing Retinal Fundus Image Classification Through Active Gradient Deep Convolutional Neural Network and Red Spider Optimization." Neural Computing and Applications 36, no. 26 (2024): 16607-16619.

[17] Mohan, Janani Priya, and Yamuna Govindarajan. "Wild Horse Optimization and Deep Learning Based Computer Aided Diagnostic Tool for Retinal Diseases." OPSEARCH (2025): 1-28.

[18] Ali, Mona AS, Kishore Balasubramanian, Gayathri Devi Krishnamoorthy, Suresh Muthusamy, Santhiya Pandiyan, Hitesh Panchal, Suman Mann et al. "Classification of Glaucoma Based on Elephant-Herding Optimization Algorithm and Deep Belief Network." Electronics 11, no. 11 (2022): 1763.

[19] Ashanand, and Manpreet Kaur. "A Novel Chaotic Weighted EHO-Based Methodology for Retinal Vessel Segmentation." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 11, no. 7 (2024): 2285455.

[20] Doddi, Guna V. "Eye Diseases Classification Dataset." Kaggle Datasets (2022).

[21] Aurangzeb, Khursheed, Sheraz Aslam, Musaed Alhussein, Rizwan Ali Naqvi, Muhammad Arsalan, and Syed Irtaza Haider. "Contrast Enhancement of Fundus Images by Employing Modified PSO For Improving the Performance of Deep Learning Models." IEEE Access 9 (2021): 47930-47945.

[22] Kansal, Kajal, Tej Bahadur Chandra, and Akansha Singh. "ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning." Procedia Computer Science 235 (2024): 70-80.

[23] Naruei, Iraj, and Farshid Keynia. "Wild Horse Optimizer: A New Meta-Heuristic Algorithm for Solving Engineering Optimization Problems." Engineering with computers 38, no. Suppl 4 (2022): 3025-3056.

[24] Huong, Audrey, KimGaik Tay, KokBeng Gan, and Xavier Ngu. "A Hierarchical Optimisation Framework for Pigmented Lesion Diagnosis." CAAI Transactions on Intelligence Technology 7, no. 1 (2022): 34-45.

[25] Tadisetty, Srikanth, Ranjith Chodavarapu, Ruoming Jin, Robert J. Clements, and Minzhong Yu. "Identifying the Edges of the Optic Cup and the Optic Disc in Glaucoma Patients by Segmentation." Sensors 23, no. 10 (2023): 4668.