# Weighted PSO Ensemble using Diversity of CNN Classifiers and Color Space for Endoscopy Image Classification

Diah Arianti[1], Azizi Abdullah[2], Shahnorbanun Sahran[3], Wong Zhyqin[4]
Centre of Artificial Intelligence-Faculty of Information Science and Technology, Universiti Kebagsaan Malaysia[1,2,3]
Bangi, Malaysia 43650, Hospital-Universiti Kebangsaan Malaysia, Cheras, Malaysia[4]

***Abstract*—Endoscopic image is a manifestation of visualization technology to the human gastrointestinal tract, allowing detection of abnormalities, characterization of lesions, and guidance for therapeutic interventions. Accurate and reliable classification of endoscopy images remains challenging due to variations in image quality, diverse anatomical structures, and subtle abnormalities such as polyps and ulcers. Convolutional Neural Network (CNN) is widely used in modern medical imaging, especially for abnormality classification tasks. However, relying on a single CNN classifier limits the model's ability to capture endoscopy images' full complexity and variability. A potential solution to the problem involves employing ensemble learning, which combines multiple models to reach at a final decision. Nevertheless, this learning approach presents several challenges, notably a significant risk of data bias. This issue arises from the unequal influence of weak and strong learners in most ensemble strategies, such as standard voting, which usually depend on certain assumptions, including equal performance among the models. However, it reduces the capability towards diverse model collaboration. Therefore, this paper proposes two solutions to the problems. Firstly, we create a diverse pool of CNNs with end-to-end approach. This approach promotes model diversity and enhances confidence in making a final decision. Secondly, we propose employing Particle Swarm Optimization to enhance the weight of the members in the ensemble learner in order to create a more resilient and accurate model compared to the standard ensemble learning approach. The experiment demonstrates that the proposed ensemble model outperforms the baseline model on both the Kvasir 1 and Kvasir 2 datasets, highlighting the effectiveness of the suggested approach in integrating diverse information from the baseline model. This enhanced performance highlights the efficacy of capturing diverse information from the baseline model.**

***Keywords*—Convolution neural network; particle swarm optimization; diversity; weighted ensemble**

## I. Introduction

The role of endoscopy images in diagnosing and treating gastrointestinal diseases is crucial. They provide a visual representation of the gastrointestinal tract, enabling the identification of abnormalities, characterization of lesions, and guidance for therapeutic interventions. The precise and reliable classification of these images continues to pose a significant challenge because of variations in image quality, diverse anatomical structures, and subtle abnormalities. These challenges highlight the need for the development of advanced techniques that can enhance classification accuracy and improve the overall effectiveness of endoscopic examinations. Computer-Aided Diagnosis (CAD) systems use advanced image processing techniques along with Artificial Intelligent (AI), Machine Learning (ML) and Deep Learning (DL) to provide reliable diagnoses. One of the remaining challenges in CAD is designing a system that can deliver the best satisfactory results for the recognition or classification in the diagnostic process. Previous research has established a crucial foundation for using Support Vector Machines (SVM) in identifying illnesses such as polyps and ulcers, as well as for leveraging Convolutional Neural Networks (CNN) in detecting bleeding [1], [2], [3], [4], [5]. Previous studies have mostly used single models for analyzing endoscopy images. However, these models may not fully capture the complexity and variability of the images. To address this limitation, an effective solution is to employ ensemble learning. This approach combines multiple models to predict a common target and make conclusive decisions, effectively mitigating bias and ultimately yielding higher-quality predictions compared to using a single classifier.

Ensemble learning has emerged as a prominent research area over the past few decades. In classification tasks, ensemble learning combines the strengths of multiple classifiers to improve overall performance by leveraging their diversity. The learning scheme distinguishes between two classifier concepts: strong and weak learner. Strong learners typically yield a lower error rate compared to weak learner, whereas weak learner predictions outperform random guessing [6]. The concept of weak and strong learners has developed into a more advanced ensemble approach known as Adaboost. This approach has further established the principles of bootstrapping and stacking. However, in the field of ML today, ensemble learning primarily revolves around bagging, boosting, and stacking. Bagging, which was originally proposed [7], significantly enhances the performance of models by applying a parallel sampling scheme to the dataset. Boosting methods, introduced in [8], train each subsequent model to rectify the mistakes made by the previous one. The classifiers in boosting are interdependent and rely on each other, leading to a collaborative learning process. The errors made by one classifier directly impact the performance of the next classifier. On the other hand, stacking involves utilizing a training model to combine predictions from multiple base learners in a diverse manner. By leveraging both base and meta learners, stacking offers a robust framework. Nevertheless, ensemble learning in ML is distinct from DL.

The predominant approach in DL research is the utilization of DL models to construct diverse model structures. This necessitates meticulous consideration of hyperparameter values, a process that demands significant time. To overcome this issue, researchers have started exploring automated hy-

perparameter optimization techniques in DL. This technique commonly involving the use of algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithms to explore the hyperparameter space and find the best combination for better model performance [9], [10]. However, automatic tuning of huge number of parameters in CNN is costly. On the other hand, fine-tuning with pre-trained weights, such as ImageNet, has been questioned in some research because it was found that the performance did not surpass that of random initialization [11], [12]. In addition, combining the DL model with different ML algorithms generates fair performance [13], [14]. Therefore, ensemble learning in DL is often presented in a simple ensemble structure, such as a standard average and majority voting approach. In the other hand, combining multiple models using those voting methods raises challenges, such as the increased risk of bias, especially when the baseline models have imbalanced performance [15].

As the remedy to the earlier problems mentioned above, in this paper primarily focuses on investigating a deep ensemble learning mechanism that underscores two crucial aspects essential for achieving successful ensemble performance: diversity and quality. As an initial step, we establish a diverse pool of Convolutional Neural Networks (CNNs). Our proposal focuses on two key aspects: (a) color transformation, and (b) model component. Furthermore, rather than relying on the average vote of the final decision among the models, we utilize Particle Swarm Optimization (PSO) to calculate the optimal weight for each individual model. This approach amplifies the strength of the more capable learners, resulting in a more resilient and precise outcome. Our main goal in this paper is to create the best combinations of models to effectively handle the wide range of variations and subtle abnormalities in the Kvasir dataset [16]. We are highly motivated to conduct this research as it is crucial for our future objectives, particularly in regard to handling actual hospital data that encompasses a diverse range of disease categories and types. Moreover, this research has the potential to assist researchers and practitioners in developing even more effective algorithms for diagnosing gastrointestinal conditions in patients with different disease categories and varying disease conditions.

In this paper, we have dedicated Section II to comprehensively discuss the existing research pertaining to our study. Building upon this literature review, we will elaborate on the methodology that we are proposing in Section III. Then, in the Section IV, we will present details about the dataset utilized, followed by a thorough discussion of the experimental results in Section V, and finally, a comprehensive conclusion in Section VI.

## II. Related Works

In a previous study by [17], researchers implemented an ensemble scheme to detect ulcers in endoscopy images. The scheme integrated multiple models such as KNN, MLP, and SVM, using a majority voting approach. The final test conducted on various color space input images demonstrated the superiority of RGB color over HIS and YCrCb, achieving an impressive accuracy of approximately 91.25%. Despite these advancements, the proposed approach has raised concerns about overfitting due to its minimal training data requirement. While, in [18], an automatic detection method for cervical precancer screening was introduced, using a larger total number of images. The authors propose a combination of three DL architectures: RetinaNet, Deep SVDD, and CNN. The ensemble model outperforms individual models in terms of performance; however, the accuracy of the results is compromised due to the presence of image noise, such as blurring. Based on the literature provided above, the use of ensemble learning extends far beyond the mere modeling of classifiers. It encompasses a crucial aspect, which is the preparation of the data prior to its input into the model. From this, it is evident that the differentiating factors among various ensemble methods revolve around the construction of the model and the fusion of the ultimate decisions. Consequently, it is important to explore two principal aspects: (i) the diversity of models, and (ii) the quality of models.

### A. Diversity

Model diversity refers to the process of generating multiple classifiers in order to introduce variation in the decision-making of the classifiers. Most of research aims to promote diversity by modifying various aspects of the network architecture, such as pre-processing data, tuning hyperparameters, and initializing weights. Data preprocessing is essential for successful classification. For example, as shown in [2], converting image data to the HSV color space and segmenting affected areas is essential for defining ulcer boundaries, thus enhancing the model's identification process. Utilizing different color spaces, such as YCbCr [3], has notably improved curvature identification accuracy using MLP. Additionally, incorporating hybrid techniques like CLAHE and Retinex can enhance polyp detection by preserving important elements like edges and texture [4], [19]. While in study [2] the utilization of filters such as Log Gabor and SUSAN corner detection greatly enhances the precision in identifying polyp boundaries using SVM. Further, CNN has numerous hyperparameters, including layer type, number of feature maps, number of neurons, kernel size, and weight. Automatically tuning all these hyperparameters with an optimization algorithm can be both costly and time-consuming. Therefore, narrowing the focus of evaluation to specific parameter such as weight can result in cost reductions.

Weight initialization in DL can be done in two main ways. The first way is by using a random distribution. The second way is by using a data-driven process. Using random initialization methods based on the Gaussian distribution can lead to slow convergence and saturated activations. To address the mentioned issue, [20] introduces an alternative approach to the one proposed by 'He' [21]. 'Glorot' assumes linear activations, while 'He' uses the ReLU activation function to introduce non-linearity in hidden layers, making 'He' initialization superior to 'Glorot' in certain DL models. Meanwhile, another Gaussian-based filter, Gabor, is a highly effective technique used to detect edges and textures in endoscopy images [2]. The Gabor filter breaks down images into different scales and orientations, allowing for a more accurate analysis of texture patterns. The Eq. (1) showcases the complex form of the Gabor filter, underscoring its intricate yet powerful capabilities.

$$G(x, y, \sigma, \theta, \lambda, \gamma, \Psi) = (-\frac{x'^2 + y^2 y'^2}{2\sigma^2})exp(i(2\Pi\frac{x'}{\lambda}) + \psi)$$
$$x' = ccos\theta + ysin\theta$$
$$y' = -csin\theta + ycos\theta \quad (1)$$

In contrast, the data-driven approach, such as [22], builds a set of patches using training images to construct new weights. The weight in this network was generated using PCA filters. The process has three stages. The first and second stages involve PCA convolution over the image patches. The third stage is the output layer, which includes data processing components like binary hashing and block-wise histogram. This filter extracts distinctive features by generating various textures for different datasets. Eq. (2) explains the process of generating PCA filters from image patches. The size of the patch in the first stage is represented by k1 and k2.

$$W_l^n = f_{k1k2}(q_1 X X^T) \in R^{k1k2}, l = 1, 2, ..., L_i \quad (2)$$

$$g_i = \left[Bhist(T_i^1), ..., Bhist(T_i^{L1})\right]^T \in R^{(2^{L2})L_1 B} \quad (3)$$

$f_{k1k2}$ is a function that maps patches to the matrix W, which will then be multiplied by the principal eigenvector $XX^T$ . While in the second stage, it is repeating the same process as in the first stage. Further, in the final stage encodes the $L_i$ images into histogram values in each block and combines them into one vector using Eq. (3). $T_i$ is the feature of the input image, while B is defined as blocks, and then the histogram of decimal value is denoted with $2^{L_2}$ . The PCA filter, however, requires $k1k2 \geq L_1, L_2$ .

### B. Quality

In terms of quality, it refers to consolidating the variance of all individual decisions. Many strategies for combining votes depend on a basic average, known as a standard method. The average vote suffers from a major drawback in making accurate predictions due to its strong bias towards weak learners [15]. While, another approach is majority voting [14], [23], [24], which collects predictions for each class label and selects the one with the highest number of votes. However, this computation becomes expensive in larger ensemble schemes and irrelevant in low-variance individual model decisions. In response to the above drawback, a weighted ensemble technique was introduced in some research. In weighted ensemble, when evaluating the weight by using validation accuracy as a metric yields comparable results to the average-based method. This is especially true when the learners demonstrate similar or slightly varied levels of accuracy. In study [25] the use of exponential function aiming for higher accuracy, however, finding the most suitable function for particular dataset for the optimal solution can be a challenging task. Therefore, in [26], different weights are automatically assigned to the learners, reflecting the unique contribution of each learner to the prediction. This method has the advantage of automatic adaptation to the new database.

## III. Methodology

In this paper, we present an ensemble learning approach that focuses on two key elements mentioned above: diversity and quality. Fig. 1 illustrates the five main processes of the proposed methodology, with detailed information as follows:

- Color-based transformation and cluster intensity – In previous work, different color space transformations were used to create sub-features for different illness categories and variation in endoscopy images. Further details in Section III(A).
- Heterogeneous network – To enhance the CNN model's extraction results, it is crucial to increase the variety of parameters and architectures utilized. Further details in Section III(B).
- Heterogeneous weight initializer – In addition to implementing various CNN architectures, it is important to utilize a range of weight initializers, such as He, Gabor, and PCA, to optimize the extraction of edges and textures within images. Further details in Section III(C).
- Classifier amplification – In the final phase, an optimized weighting was proposed to quantify the strong classifier's contribution within the ensemble. Further details in Section III(D).

### A. Color-based Transformation and Cluster Intensity

Endoscopy images commonly utilize the RGB color channel representation. However, other well-known color channel representations, such as HSV, CIE-LAB, and YCrCb, are frequently employed in diverse medical image analyses. Various representations reveal abnormal patterns, such as color and geometric characteristics observed in cases of polyps and ulcers [2], [4], [13]. Drawing inspiration from the image capturing procedure [27], where the light source moves along one side of the narrow path within the GI tract, we make the assumption that objects closer to the light source tend to have higher luminance. Considering this, we propose three distinct region to tackle the complexity of intensity variation in image samples:

- The outer area (C1) – This region offers the most intense illumination. This area is designed to clearly identify any protruding objects in this region, such as polyps and folds in the colon.
- The inner area (C2) - This area is adjacent to the '*outer*' area. In this region, the blue area that surrounds the polyps in the 'dyed' category is expected to be distinct from the protruding part of the polyp.
- The junction area (C3) - This region is the farthest area from the source of light. We aim to identify the shared characteristics of renowned landmarks in this particular region, including the cecum, pylorus, and z-line.

We apply the k-means algorithm, with a maximum of 3 clusters, to determine the optimal solution for the given assumptions above. For this clustering process, we consider the gray image in RGB, the 'L' or luminance component in LAB,

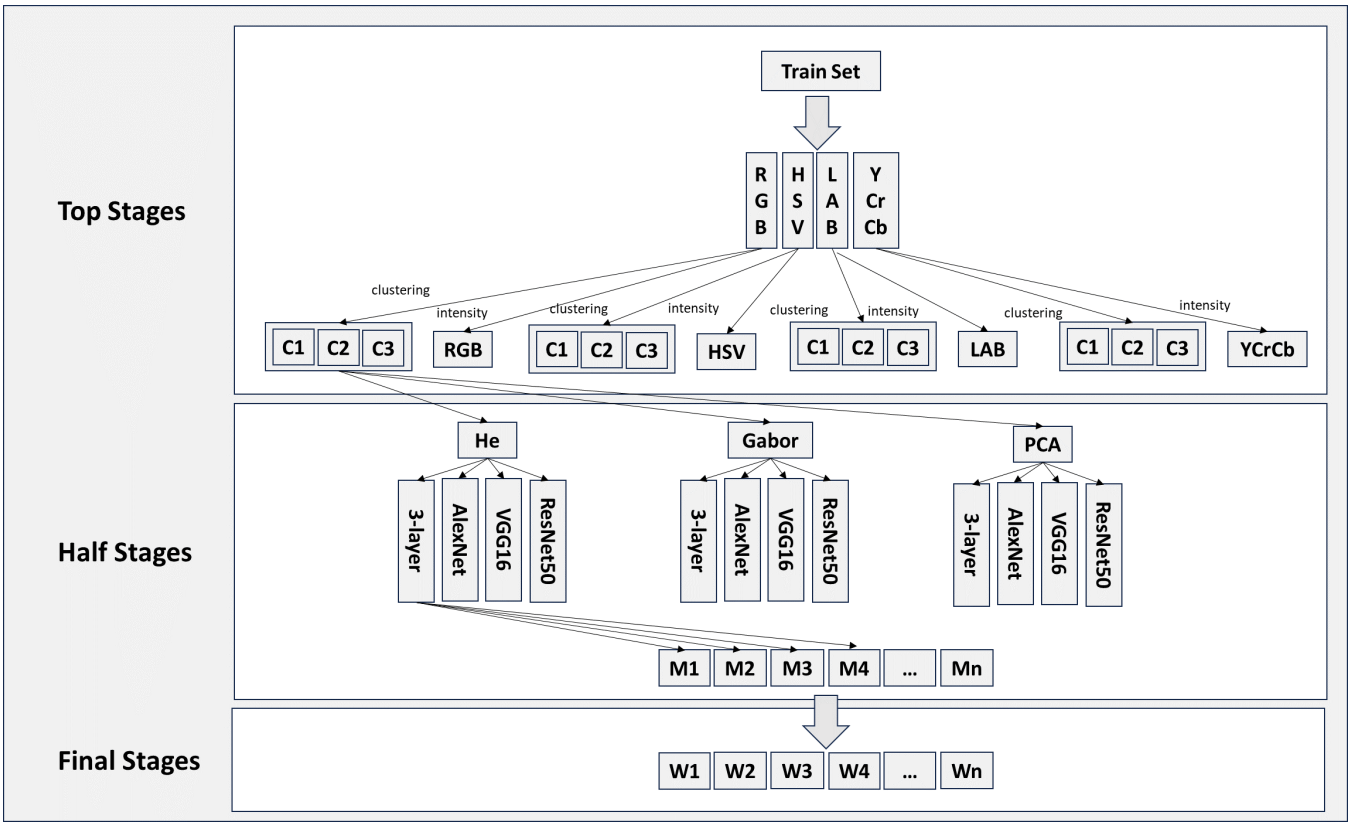


Fig. 1. The proposed ensemble architecture. The top stages focusing on preprocessing data, then in half stages focus on creating diverse CNN pool and the final stage focus more on classifier amplification.

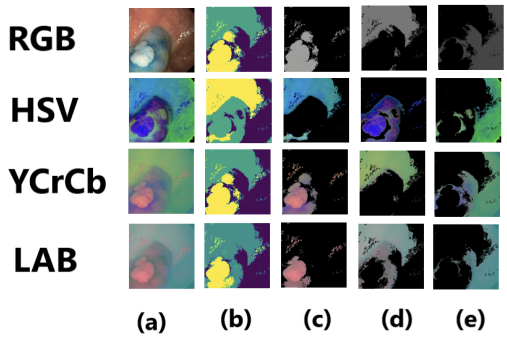


Fig. 2. Clustering of three distinct areas in the 'dyed lifted polyp' category: (a) Original image, (b) K-means cluster, (c) C1, (d) C2, (e) C3.

the 'Y' or luminance component in YCrCb, as well as the 'V' or value component in HSV. We use various intensity schemes to accurately represent tissue colors and ensure robustness to lighting variations. Fig. 2 displays an image transformation of four color channels from the Kvasir dataset in the 'dyed-lifted polyp' category. Column A shows the original images, while column B displays the three clusters obtained through the application of k-means clustering. Next, in column C, the outer area is revealed after the mask is applied to the original image. Column D showcases the inner area, followed by the junction area in column E.

### B. Heterogenous Network

Convolutional Neural Network (CNN) is widely regarded as one of the most popular deep neural network models. It is composed of powerful components such as convolutional layers, pooling layers, and non-linear activation functions [28]. In this paper investigates four different CNN architectures in the context of the study: a 3-layer CNN, AlexNet [28], VGG16 [29], and ResNet50 [30]. To emphasize the use of a shallow CNN architecture (3-layer network) in our proposed network, we have incorporated the branch CNN concept from [31]. The Branch CNN represents a new variation of the traditional CNN, implementing the concept of *"coarse to fine"* by establishing a separate branch on VGG16. One of the crucial features of this architecture is the inclusion of a weight in the loss function, which ensures a precise representation of the branch's influence on the overall loss.

$$L_j = \sum_{n=1}^{N} -W_n log \frac{e^{f_y^a} j}{\sum_i e^{f_i^a}} \tag{4}$$

Eq. (4) presents the entropy loss function ($L$) in conjunction with the weighted loss value. Branch implementation was not carried out in VGG16 as we had anticipated that the number of layers in AlexNet would align with those in VGG16 when incorporating branches. Furthermore, we are reducing the number of layers in AlexNet into 3-layer, while evaluating their potential to deliver equivalent performance improvements.

The weight loss values in the branch were determined by conducting three separate runs. This resulted in weight loss values of 0.4 for the 3-layers and 0.6 for AlexNet. In advanced configurations, the ReLU activation function is used along with the implementation of 8-fold cross-validation, ensuring a 70:30 ratio between training and validation. However, our focus lies on determining the highest level of accuracy from these variations. Furthermore, essential parameters such as the learning rate have been set at 0.001, while the optimizer follows the SGD algorithm. To mitigate the risk of overfitting, we have incorporated an early stopping mechanism, limiting the maximum epoch to 50.

### C. Heterogeneous Weight Initializer

This paper uses both the random-based and data-driven initialization techniques mentioned above, which are 'He' [12], Gabor, and PCA [22]. To create the "He" filter, we utilize existing libraries in Keras. On the other hand, for the Gabor filter, we generate a filter bank consisting of multiple filters with four different parameters ( $\sigma, \theta, \lambda, \gamma$ ). Considering that we employ a CNN consisting of multiple filters, we assume that the Gabor bank also consists of the same number of filters as the CNN. Next, PSO was utilized to obtain the parameters for Gabor filters. Inspired by [32] work, the proposed method apply SVM as an evaluator in order to find the best parameter values for each filter in the Gabor bank. Additionally, we anticipate addressing the distribution issue for each PSO particle value through a 'centroid' approach. Meanwhile, Fig. 3 exemplifies the outcomes of generating a filter using the PCANet concept on the Kvasir v1 datasetin training process. In part (a), we can observe the filter applied to VGG19, while in image (b) the filter is applied to AlexNet.

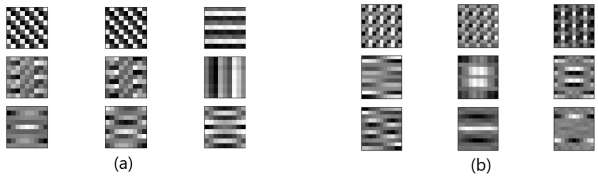


Fig. 3. Training samples of the PCANet with size 9x9 (a) VGG16 and 11x11 (b) on AlexNet using RGB-color data.

Finally, the various parameter combinations mentioned previously result in approximately 1536 models. Rather than generating the entire model as mentioned above, we opted for a simplified elimination strategy to expedite execution time. Our focus is on minimizing parameter variability in fold formation during cross-validation. This method entails selecting the best-performing model based on the average accuracy across folds. In this case, we simply choose RGB and YCrCb color transformation with the 'He' as part of this selection. The number of models in the experiment is reduced by approximately 87.5%. The proposed CNN pool finally contains a total of 192 baseline models.

### D. Classifier Amplification

Assuming a balanced performance across all learners, it is essential to assign equal weight to each classifier, similar to the standard average ensemble scheme. However, the diverse concepts in ensemble learning can lead to imbalanced performance, which can ultimately affect the overall performance. Thus, in the proposed approach, we utilize a swarm-based optimization algorithm, PSO to fine-tune the weight and achieve a balanced outcome. In [Eberhart, 1995] introduced PSO, a population-based evolutionary computational algorithm that solves optimization problems involving a lack of domain knowledge. The population is like a flock of birds that can maintain individual position and speed while flying in a specific direction. The standard PSO formulation is described by Eq. (5)

$$V_i = \omega V_i + c_1 r_1 (Pbest - X_i) + c_2 r_2 (Gbest - X_i) V_{i+1} = X_i + V_i \quad (5)$$

$V_i$ and $X_i$ represent the velocity and current position of particle i, while Pbest is the best personal position and Gbest is the best global position for all particles in the population. $\omega$ represents the inertia weight. c1 and c2 are the acceleration coefficients that improve the exploitation ability of each particle. r1 and r2 are the random numbers that increase exploration ability. PSO is more focused on searching for values in space based on velocity, in contrast to other optimization algorithms like Genetic Algorithm (GA). Therefore, PSO is suitable for continuous-valued problems and enables faster convergence [32]. Fig. 5 illustrates the proposed algorithm for fine-tuning the weights. In the beginning, it involves 20 particles, each representing an individual agent within the ensemble. These agents collaborate to discover the weight combination that delivers the highest accuracy performance. The accuracy of the ensemble, using the optimal weight from the best personal weight (pBest), is employed to compute f. If the pBest outperforms the current best global weight (gBest), then gBest is updated with the new pBest. Throughout the 100 epochs, the agent with the highest final accuracy in gBest is deemed the top agent. Eq. (6) is employed to initialize the PSO inertia weight.

$$\omega_i = \omega_{max} - \left(\frac{\omega_{max} - \omega_{min}}{max_{iter}}\right) * i \quad (6)$$

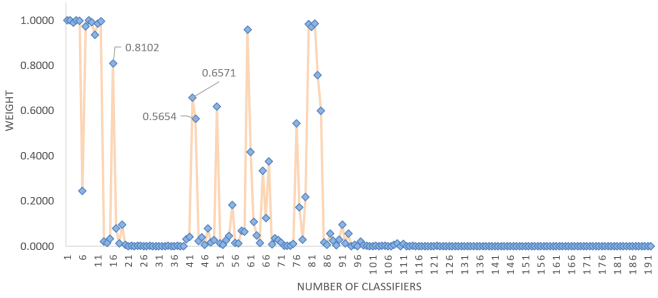


Fig. 4. The optimum PSO weight on the best performance of the proposed method.

## IV. Dataset and Metrics

### A. Dataset

We used the Kvasir dataset [16], which contains images of patients' upper and lower gastrointestinal tract including normal and pathological findings such as polyps and ulcers.

There are two versions of the Kvasir dataset, which are v1 and v2. The first version has 500 images in each class, with eight total classes. Thus, we have 4000 images. The dataset's original resolution varied from 720 × 576 to 1920 × 1072. Then, it is cropped and resized to a resolution of 227 x 227. While the second version contains 1000 images per category, half were available in the first version. Thus, since we use the first version for training, then in total, we have new test data from Kvasir v2, which is about 4000 total images. As in the training process, we use 80% of Kvasir v1. To reduce the risk of overfitting in our baseline model, we apply data augmentation techniques such as zoom, shear, rotate, and width shifting to the training set. Thus, after augmentation, the total training dataset contains 16,800 images. During testing, we used two datasets: Kvasir v1 with 800 images and Kvasir v2 with 4000 images.

### B. Metrics

This paper uses performance metrics such as accuracy, precision, sensitivity, and specificity to evaluate our baseline and proposed method. Accuracy describes the ability of the model to detect the correct classes in this classification as shown in Eq. (7):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

TP (True Positive) is when the input is correctly predicted as positive. TN (True Negative) is when the input is correctly predicted as negative. FP (False Positive) is when the input is incorrectly predicted as positive. FN (False Negative) is when the input is incorrectly predicted as negative. In contrast, sensitivity is the ratio of true positives found to positives in the dataset as shown in Eq. (8).

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

Specificity is the ratio of true negatives found to negatives in the dataset as shown in Eq. (9).

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

Afterwards, the precision of the prediction can be measured by calculating the amount of positive predictive data, as indicated in Eq. (10).

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

## V. Experimental Result

This section categorizes the experiment into two main objectives based on the insights gained from developing a solution. First, it delves into analyzing the impact diversity concept in ensemble learning. Second, identifying strong and weak learners is important for dealing with the significant impact of weight amplification in our system. This approach allows us to clearly observe the most influential learners in the proposed method.

### A. Diversity Impact in Proposed Scheme

In this context, three main group experiments need to be conducted. The first group comprises 12 models that utilize the RGB color space without clustering (see Table I-c and f). This group was formed to assess the impact of the classifiers, particularly those that demonstrate the higher accuracy within the pool. Next, in the second group, 48 models from 4 different color spaces (RGB, YCrCb, HSV, and LAB) without clustering were created to determine the comparison of contributions with the first group (see Table I (d) and (g)). The performance of groups 1 and 2 was significantly different, particularly regarding the RGB and non-RGB input models. Then the third group contains all the proposed models in the pool, namely 192 models (see Table I (e) and (h)).

In standard approach with Kvasir v1, the top-1 baseline accuracy for the third group was achieved using AlexNet and Gabor filter with RGB color space, without clustering. This single model accuracy, in training, reaches about 88.7% and in testing it achieved 84% accuracy. In this scenario, the standard ensemble enhances accuracy by approximately 1%. Further, by focusing only on the first group, the risk of overfitting was minimized. Further, the issue was resolved using the second group test. In this case, the accuracy of the model increased by approximately 6% compared to the baseline. While in the proposed method, experiments in the first group yielded significant results, as did those in groups two and three. As diversity increases, accuracy also increases. This demonstrates the presence of distinct features within each existing group. Interestingly, this stands in contrast to the performance of the standard method. However, in reality, maintaining a balance of performance among a collection of models can be quite challenging, particularly if the goal is to decrease the execution time for generating models. Based on the significant differences in tests in first groups (average accuracy is 81.9%) and second groups (average accuracy is 68.8%), there is no guarantee that maintaining balance in overall model performance will result in improved performance. This is a notable weakness of the standard ensemble model. Additionally, the data from Kvasir v2 showcases that experiments involving all three groups consistently show instability in standard approach, especially when compared to the proposed method.

### B. Strong and Weak Learner

In this scenario, the experiment was conducted 50 times, yielding results that demonstrate the tremendous potential of the proposed method in enhancing the accuracy of the maximum single model. Specifically, our findings reveal an improvement of 7%. (see Table I (b) and (h)). Although the model has imbalanced performance, the accuracy was improved after classifier amplification. Fig. 4 shows the optimum weight value after amplification on the training dataset. The strong learner is identified by a weight greater than the mean, while the weak learner is identified by a weight smaller than the mean. According to the data in Fig. 4, there are 30 strong learners in this group, with most of the top 15 strong learners coming from the RGB color space without clustering. However, there are three learners among the strong learners coming from YCrCb intensity clustering in C1 and C3 (it was labelled in Fig. 5 – C1 is 0.8102 and C3 with 0.6571 and 0.5654). Furthermore, the experiment was continued excluding

TABLE I. THE PERFORMANCE COMPARISON OF THE PROPOSED SCHEME AND STANDARD SCHEME ON TEST DATA

| Dataset | Metrics | Baseline | | Standard Average | | | Proposed | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (a) Average | (b) Maximum | (c) Full-RGB(12) | (d) Full-All Intensity(48) | (e) All model(192) | (f) Full-RGB(12) | (g) Full-All Intensity(48) | (h) All model(192) |
| Kvasir v1 | Accuracy | 0.6050 | 0.8400 | 0.8788 | 0.9000 | 0.8500 | 0.8900 | 0.8988 | 0.9100 |
| | Precision | 0.5975 | 0.8474 | 0.8820 | 0.9008 | 0.8576 | 0.8927 | 0.9000 | 0.9108 |
| | Sensitivity | 0.6057 | 0.8400 | 0.8788 | 0.9000 | 0.8500 | 0.8900 | 0.8988 | 0.9100 |
| | Specificity | 0.6063 | 0.8401 | 0.8788 | 0.9000 | 0.8502 | 0.8901 | 0.8988 | 0.9100 |
| Kvasir v2 | Accuracy | 0.8530 | 0.8918 | 0.9153 | 0.8540 | 0.8760 | 0.9143 | 0.9098 | 0.9158 |
| | Precision | 0.8614 | 0.8964 | 0.9179 | 0.8821 | 0.8882 | 0.9169 | 0.9175 | 0.9228 |
| | Sensitivity | 0.8530 | 0.8918 | 0.9153 | 0.8540 | 0.8760 | 0.9143 | 0.9098 | 0.9158 |
| | Specificity | 0.8532 | 0.8918 | 0.9153 | 0.8545 | 0.8763 | 0.9143 | 0.9099 | 0.9159 |

weak learners, and achieved the identical level of accuracy as when weak learners were involved, specifically 91%. However, performance decreased when tuning the weights for the 30 classifiers mentioned earlier, dropping by around 0.875%. Additionally, Table II demonstrates the impact of the proposed clustering method on various color spaces by showing the most influential learner based on color intensity clustering after 50 runs. It indicates that YCrCb is the only color space that influences significantly the performance of the proposed method. While the other learner still makes a contribution, their impact on the proposed scheme's performance is very limited.

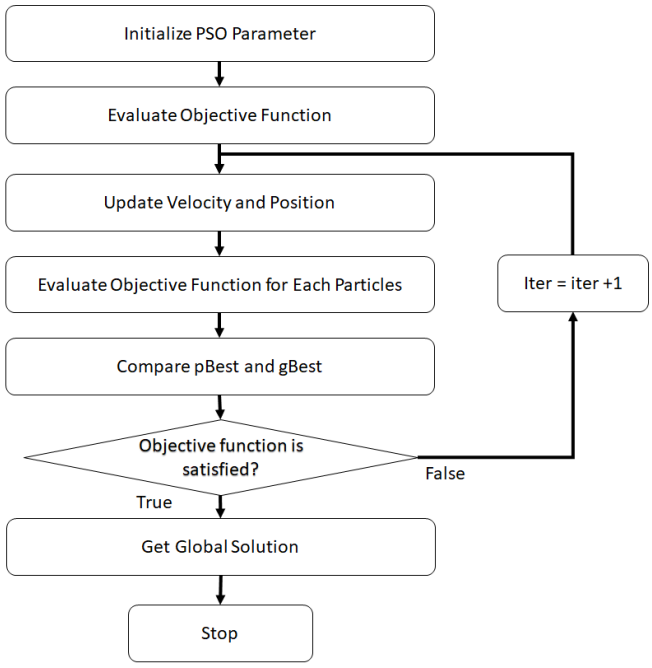


Fig. 5. The computational algorithm of PSO.

TABLE II. MOST SIGNIFICANT LEARNER IN THE GROUP OF COLOR CLUSTERING

| Rank | Network | Filter | Color | Area | n-runs |
|---|---|---|---|---|---|
| 1 | VGG16 | He | YCrCb | Outer | 50 |
| 2 | AlexNet | Gabor | YCrCb | Junction | 8 |
| 3 | ResNet50 | PCA | YCrCb | Junction | 7 |
| 4 | ResNet50 | Gabor | YCrCb | Junction | 2 |
| 5 | AlexNet | He | YCrCb | Junction | 1 |

## VI. CONCLUSION

This study introduces a CNN-based ensemble method designed to enhance the accuracy of classifying the Kvasir dataset. The experimental results demonstrate that the proposed method surpasses the standard approach by delivering consistent performance across diverse test datasets. The utilization of color intensity-based clustering prioritizes notable features, particularly in abnormal cases such as polyps, ulcers, esophagitis, and "dyed" categories. By employing various CNN hyperparameters to create a range of models in the ensemble, the risk of overfitting is reduced in both the standard and proposed methods. This approach not only enhances the learning process but also unveils the potential of features in various color space transformations and color intensity-based clustering.

In conclusion, we can summarize the findings and drawbacks as follows: Firstly, the Kvasir dataset displays unique characteristics when the data is converted into different color spaces, such as RGB, HSV, YCrCb, and LAB. Secondly, clustering a specific region within the image, specifically related to conditions like polyps and ulcers, leads to diverse responses and significantly impacts the overall performance of the model, particularly in the YCrCb color space. However, this imbalance in the overall model performance hinders the attainment of a high standard ensemble accuracy. Even after trying different pre-processing methods, the accuracy is still consistently lower compared to datasets that are not clustered. Moreover, it is essential to enhance the diversity of models in order to achieve optimal results with the proposed method. Simultaneously, by amplifying the learner, we can effectively mitigate the risk of overfitting in the standard scheme. It is important to note that both mechanisms are crucial for improving the scheme's overall performance. Moreover, the 3-layer network architecture is an integral part of AlexNet and incorporates the concept of branch CNN. Conducting experiments with other network types could potentially yield significant advantages in addressing the limitations of the proposed method, such as applying it to VGG16. Furthermore, we recommend utilizing alternative search algorithms, such as genetic algorithms or the Bee's algorithm, to boost mutation capacity during the training phase and decrease the execution time required to generate Gabor banks.

## ACKNOWLEDGMENT

This work has been supported by the University Kebangsaan Malaysia Research Grant GUP-2021-063 and

Malaysia's Ministry of Higher Education Fundamental Research Grant FRGS/1/2019/ICT02/UKM/02/8

## REFERENCES

[1] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. A. G. Johansen, J. Rittscher, M. A. Riegler, and P. L. Halvorsen, "Real-Time Polyp Detection , Localization and Segmentation in Colonoscopy Using Deep Learning," vol. 9, 2021.

[2] A. Karargyris and N. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos." *IEEE transactions on biomedical engineering*, vol. 58, no. 10, pp. 2777–2786, 10 2011.

[3] B. Li, L. Qi, M. Q. Meng, and Y. Fan, "Using ensemble classifier for small bowel ulcer detection in wireless capsule endoscopy images," in *2009 IEEE International Conference on Robotics and Biomimetics, ROBIO 2009*, 2009, pp. 2326–2331.

[4] V. Vani and K. V. M. Prashanth, "Ulcer detection in Wireless Capsule Endoscopy images using deep CNN," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3319–3331, 2022.

[5] M. Hajabdollahi, R. Esfandiarpoor, P. Khadivi, S. M. R. Soroushmehr, and N. Karimi, "Biomedical Signal Processing and Control Segmentation of bleeding regions in wireless capsule endoscopy for detection of informative frames," *Biomedical Signal Processing and Control*, vol. 53, p. 101565, 2019.

[6] G. Stavropoulos, R. V. Voorstenbosch, F.-j. V. Schooten, and A. Smolinska, *Random Forest and Ensemble Methods*, 2nd ed. Elsevier Inc., 2020.

[7] Leo Breiman, "Bagging Predictors," vol. 140, pp. 123–140, 1996.

[8] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[9] E. Ayan, H. Erbay, and F. Varc¸ın, "Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 179, 2020.

[10] Y. Wang, H. Zhang, and G. Zhang, "cPSO-CNN: An efficient PSObased algorithm for fine-tuning hyper-parameters of convolutional neural networks," *Swarm and Evolutionary Computation*, vol. 49, pp. 114–123, 2019.

[11] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Feedforward neural networks initialization based on discriminant learning," *Neural Networks*, vol. 146, pp. 220–229, 2022.

[12] K. He, R. Girshick, and P. Dollar, "Rethinking imageNet pre-training," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, no. ii, 2019, pp. 4917–4926.

[13] M. M. Rahman, M. A. H. Wadud, and M. M. Hasan, "Computerized classification of gastrointestinal polyps using stacking ensemble of convolutional neural network," *Informatics in Medicine Unlocked*, vol. 24, p. 100603, 2021.

[14] B. Zhang, S. Qi, P. Monkam, C. Li, F. Yang, Y. D. Yao, and W. Qian, "Ensemble learners of multiple deep cnns for pulmonary nodules classification using ct images," *IEEE Access*, vol. 7, pp. 110 358–110 371, 2019.

[15] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[16] K. Pogorelov, K. Ranheim Randel, C. Griwodz, T. de Lange, V. Viken Health Trust, N. Dag Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. Thelin Schmidt Karolinska Institutet, S. Karolinska Hospital, S. Michael Riegler, P. Halvorsen, S. Losada Eskeland, D. Johansen, P. Thelin Schmidt, and M. Riegler, "Kvasir: A Multi-Class Image- Dataset for Computer Aided Gastrointestinal Disease Detection Sigrun Losada Eskeland," *ACM Reference format*, 2017.

[17] B. Li and M. Q. Meng, "Texture analysis for ulcer detection in capsule endoscopy images," *Image and Vision Computing*, vol. 27, no. 9, pp. 1336–1342, 2009.

[18] P. Guo, Z. Xue, Z. Mtema, K. Yeates, O. Ginsburg, M. Demarco, L. Rodney Long, M. Schiffman, and S. Antani, "Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening," *Diagnostics*, vol. 10, no. 7, pp. 1–13, 2020.

[19] M. A. Khan, S. Kadry, M. Alhaisoni, Y. Nam, Y. Zhang, V. Rajinikanth, and M. S. Sarfraz, "Computer-Aided Gastrointestinal Diseases Analysis from Wireless Capsule Endoscopy: A Framework of Best Features Selection," *IEEE Access*, vol. 8, pp. 132 850–132 859, 2020.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Journal of Machine Learning Research*, vol. 9, 2010, pp. 249–256.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, 2015, pp. 1026–1034.

[22] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?" IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5017–5032, 2015.

[23] D. Albashish, S. Sahran, A. Abdullah, M. Alweshah, and A. Adam, "A hierarchical classifier for multiclass prostate histopathology image gleason grading," *Journal of Information and Communication Technology*, vol. 17, no. 2, pp. 323–346, 2018.

[24] T. Roß, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. Mindroc-Filimon, P. Scholz, T. N. Tran, P. Bruno, P. Arbel´aez, G. B. Bian, S. Bodenstedt, J. L. Bolmgren, L. Bravo-S´anchez, H. B. Chen, C. Gonz´alez, D. Guo, P. Halvorsen, P. A. Heng, E. Hosgor, Z. G. Hou, F. Isensee, D. Jha, T. Jiang, Y. Jin, K. Kirtac, S. Kletz, S. Leger, Z. Li, K. H. Maier-Hein, Z. L. Ni, M. A. Riegler, K. Schoeffmann, R. Shi, S. Speidel, M. Stenzel, I. Twick, G. Wang, J. Wang, L. Wang, L. Wang, Y. Zhang, Y. J. Zhou, L. Zhu, M. Wiesenfarth, A. Kopp-Schneider, B. P. M¨uller-Stich, and L. Maier-Hein, "Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge," *Medical Image Analysis*, vol. 70, p. 101920, 2021.

[25] B. K. Kim, J. Roh, S. Y. Dong, and S. Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[26] A. Qasem, S. Sahran, S. N. H. S. Abdullah, D. Albashish, R. I. Hussain, and S. Arasaratnam, "Heterogeneous ensemble pruning based on Bee Algorithm for mammogram classification," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, pp. 231–239, 2018.

[27] J. L. Buxbaum, D. Hormozdi, M. Dinis-Ribeiro, C. Lane, D. Dias-Silva, A. Sahakian, P. Jayaram, P. Pimentel-Nunes, D. Shue, M. Pepper, D. Cho, and L. Laine, "Narrow-band imaging versus white light versus mapping biopsy for gastric intestinal metaplasia: a prospective blinded trial," in *Gastrointestinal Endoscopy*, vol. 86, no. 5. Elsevier Inc., 2017, pp. 857–865.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Handbook of approximation algorithms and metaheuristics," pp. 1–1432, 2007.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–14.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016.

[31] X. Zhu and M. Bain, "B-CNN: Branch Convolutional Neural Network for Hierarchical Classification," 2017.

[32] S. Khan, M. Hussain, H. Aboalsamh, H. Mathkour, G. Bebis, and M. Zakariah, "Optimized Gabor features for mass classification in mammography," *Applied Soft Computing Journal*, pp. 1–14, 2016.