Journal of Computer Science and Cybernetics, V.41, N.xx (2025), 1–
DOI: 10.15625/1813-9663/22535

A NOVEL WEIGHTED ENSEMBLE APPROACH FOR ENHANCING IMAGE RETRIEVAL EFFECTIVENESS WITH DEEP LEARNING MODELS

TRAN VAN KHANH[1,2], NGUYEN NGOC THUY[1], NGO MINH HUONG[3], LE MANH THANH[1,*]

[1] *University of Sciences, Hue University, 77 Nguyen Hue, Hue City, Viet Nam*

[2] *University of Khanh Hoa, 01 Nguyen Chanh Street, Nha Trang Ward, Khanh Hoa Province, Viet Nam*

[3] *Institute for Artificial Intelligence, University of Engineering and Technology, VNU, 144 Xuan Thuy Street, Cau Giay Ward, Ha Noi, Viet Nam*

Abstract. Content-based image retrieval (CBIR) is becoming increasingly important amid the rapid growth of image data. Traditional CBIR approaches, which rely on features such as color, shape, and texture, often face limitations in accuracy. Even when using features extracted from deep learning models, these approaches still fall short of fully meeting user expectations. To enhance retrieval effectiveness, this study introduces an ensemble approach that utilizes feature sets from multiple deep learning models. In our method, retrieved images are determined through an aggregation of recommendations from deep learning models, with each model's vote assigned a specific weight. This weight is comprehensively evaluated based on the similarity between the recommended image and the query, the model's reliability, and the distribution of images recommended by each model. To validate the effectiveness of this approach, we conducted experiments using VGG16, ResNet50, EfficientNetB0, DenseNet201, Swin, and Clip, pre-trained on ImageNet for feature extraction. Three model combinations, (1) VGG16, ResNet50, and EfficientNetB0, (2) ResNet50, EfficientNetB0, and DenseNet201, and (3) Swin and Clip, were explored within the proposed ensemble framework on the Oxford-17-Flowers, Caltech-101, CIFAR-10, and ISIC-2018 datasets. The suggested approach routinely outperforms individual models, according to experimental results, providing better retrieval accuracy on most datasets.

Keywords. CBIR, Ensemble approach, VGG16, ResNet50, EfficientNetB0, DenseNet201.

1. INTRODUCTION

In the current digital era, the volume of digital image data generated and utilized by users has increased significantly. The necessity for image retrieval has become intense and is a significant challenge, representing a critical issue in the domain of computer vision. CBIR requires finding images

*Corresponding author.
E-mail addresses: tvkhanh.dhkh24@hueuni.edu.vn (T.V. Khanh); nnthuy.cs@hueuni.edu.vn (N.N. Thuy); nmhuong@vnu.edu.vn (N.M. Huong); lmthanh@hueuni.edu.vn (L.M. Thanh).

that exhibit similar content to a specified input image from an extensive dataset [1]. CBIR is an entirely automated procedure that entails feature extraction from the query image and the primary dataset [2]. Feature extraction is the initial phase of CBIR, employed to convert the abstract content of an image into numerical representations that machines can process [3]. Low-level features characterize the image, but high-level features encapsulate the perceptual information included within the image. The disparity between low-level and high-level features is termed the semantic gap [4]. The precision of CBIR is significantly affected by the feature extraction method. Conventional feature extraction techniques frequently depend on manually generated attributes such as hue, form, and texture. Nonetheless, these methods encounter considerable constraints concerning efficiency, accuracy, and the capacity to depict intricate images [5, 6, 7, 8]

The advent and progression of deep learning models have introduced a novel trajectory for CBIR. Deep learning models, especially Convolutional Neural Networks, possess the capability to autonomously learn and extract features from intricate images, hence enhancing search performance and accuracy [9]. These models not only acquire robust feature representations but also possess the potential to scale to extensive datasets. The current research trend emphasizes deep learning models for feature extraction, hence improving the efficacy of image retrieval systems [10, 11, 12, 13, 14].

Deep learning models yield superior outcomes to traditional image feature extraction techniques. However, deep models necessitate substantial data for learning and weight optimization. Insufficient training data may result in these models experiencing overfitting or underfitting, leading to erroneous feature extraction outcomes [15]. Furthermore, the capacity of transfer learning across domains is a significant concern in machine learning and deep learning when a model trained on one dataset must be utilized on another with distinct properties. Domain shifts can impair the performance of machine learning models, as the features acquired during training may be ineffective in the target domain. Correspondingly, feature fusion techniques in CBIR problems also frequently yield enhanced performance. Nonetheless, it presents numerous challenges, including increased system complexity, which leads to greater resource consumption in terms of time and memory, and compatibility issues among features from different models.

Despite recent progress in CBIR, several limitations persist. Many traditional approaches still rely on handcrafted or shallow features, which lack the ability to represent high-level semantic information. Although deep learning methods have achieved significant improvements, most studies are restricted to a single architecture (e.g., CNN), resulting in unstable performance when handling heterogeneous datasets. Furthermore, feature fusion or ensemble strategies employed in prior work are often simplistic, typically involving fixed linear combinations that fail to consider the reliability or contextual relevance of individual models. Another important shortcoming is the limited use of statistical validation techniques, such as t-tests or confidence intervals, which raises concerns regarding the robustness and academic reliability of reported results. These gaps motivate our research and raise three central questions. First, how can features extracted from multiple deep learning models be effectively combined to enhance CBIR performance? Second, is it feasible to design a weighted ensemble mechanism that dynamically adapts to both model reliability and similarity levels? Finally, can the proposed approach achieve statistically significant improvements over individual models?

As a result, refining the feature fusion process to meet accuracy requirements for the CBIR problem across diverse datasets is crucial for improving retrieval efficiency and utilizing the advantages of various deep learning models. This work presents an integrated methodology that employs merged feature sets from various deep learning models. The suggested method entails that each resultant image amalgamates elements from various deep learning models. Each image in the aggregated

result is allocated a specific weight, which is thoroughly assessed based on the similarity between the suggested images and the query image, the dependability of the models, and the distribution of images offered by each model.

The main contribution of this study can be listed as follows:

1) Propose a method that combines multiple deep learning models to enhance the effectiveness of the CBIR task. Specifically, we introduce a weighting strategy for images in the consolidated result set, leveraging the model's reliability, the similarity between the query image and the result images, as well as the standard deviation in the related image set. This method helps optimize the combination of models, improving the overall accuracy of the CBIR system.

2) We conducted experiments on various datasets to evaluate the performance of the proposed method. Experimental results demonstrate that this combined strategy not only enhances the accuracy of image retrieval but also ensures stability and generalization across various datasets, including the three combinations *VgReEf_CBIR, ReEfDe_CBIR, and SwCl_CBIR.*

The rest of the paper is organized as follows: Section 2 presents related works; Section 3 introduces the proposed overall approach; experimental results and discussion are reviewed in Section 4, and finally, Section 5 presents the conclusions of the study.

2. RELATIVE WORKS

In this section, we will briefly present some recent related studies, specifically focusing on studies using deep learning models for feature extraction of CBIR problems.

A deep learning-based approach for CBIR on facial image data is introduced in [16]. Singh et al. proposed a deep convolutional neural network (CNN) that extracts features using the activations of convolutional layers and employs max-pooling as a feature reduction technique. Additionally, the method incorporates partial feature mapping as image descriptors to effectively capture the repetitive patterns inherent in facial images. In a related study [17], a CBIR-CSNN model was developed by integrating convolutional Siamese neural networks with CBIR frameworks. These models demonstrate strong performance in distinguishing CT images, highlighting their effectiveness in medical image retrieval tasks. Eswaran et al. [18] proposed an image search engine system that leverages various pre-trained CNN models for feature extraction. Their approach focuses on making minimal modifications to the ResNet-50 architecture to adapt it to specific datasets, thereby enhancing retrieval performance. Similarly, Reena et al. [19] utilized deep learning features extracted from the activations of the feature layers in a pre-trained ResNet-101 network. To improve efficiency, they applied t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction. For image retrieval, similarity between images is computed using Euclidean distance, enabling the identification of analogous images from a database.

Some research indicates that integrating deep learning models produces good results. Sharma et al. [20] have proposed utilizing the VGG19 and Squeeze models, which have been trained on extensive datasets for feature extraction. The two feature vectors generated by these suggested models are combined to provide a vector with enhanced features. The authors utilized a classification technique to categorize the integrated features, minimizing search times. The Euclidean distance metric quantifies the distance between the query image and the images within the dataset. The model suggested demonstrates notable performance relative to deep learning models. However, the proposed combination of features leads to large feature vectors, which affects computational cost. Besides, Kumar et al. [11] have proposed the integration of DarkNet-19 and DarkNet-53 for feature extraction. The

DarkNet-53 model employs the final convolutional layer for feature extraction, conserving spatial information and linking to all input neurons. The average pooling layer serves as the feature extraction layer in DarkNet-19. Using the PCA algorithm to reduce feature dimensionality, therefore minimizing computational expenses. However, excessive dimensionality reduction can lead to information loss when applying the PCA algorithm. Additionally, transferring to a new space via PCA could impact the relationships between components, so it fails to maintain the original feature information. Recent studies have increasingly focused on ensemble and feature fusion strategies to enhance the robustness of deep learning models. For instance, Bhandari et al. [21] introduced a weighted average ensemble model for brain tumor classification in MRI images, demonstrating that assigning weights to individual models can significantly improve predictive accuracy compared to single architectures. Similarly, Lee and Kim [22] proposed an adaptive ensemble learning approach, in which intelligent feature fusion dynamically integrates the outputs of multiple deep networks. This adaptive mechanism enables more flexible model combinations and better performance across diverse datasets. In another direction, Singh et al. [23] developed a hierarchical deep feature fusion and ensemble framework for MRI classification, which successfully leverages multi-level feature representations to boost classification accuracy. While these studies highlight the advantages of weighted ensembles and adaptive fusion particularly in improving accuracy and generalization they also present certain limitations. Most methods are tailored for domain-specific tasks such as medical imaging, and their weighting mechanisms are often predefined or lack statistical validation. Consequently, there remains a need for more generalizable and statistically grounded ensemble strategies that can be effectively applied to the CBIR problem. Additionally, Chughtai et al. [24] have introduced models including VGG16, VGG19, EfficientNet, ResNet50, and 12 CNN variations for the CBIR task utilizing transfer learning methodologies. The study looks at and finds the best hyperparameters for 16 different types of CNNs by looking at things like complexity, execution time, data preprocessing methods, and more. This research provides a comprehensive insight into the influence of hyperparameters on transfer learning methodologies.

The survey and analysis of related works, show the feasibility and effectiveness of deep learning methods in the CBIR, significantly improving performance compared to traditional techniques. However, many methods still have limitations in terms of accuracy, generalization ability, and efficiency when applied to large datasets. Based on the published works, this paper proposes a new hybrid model that integrates multiple deep learning components to address the limitations of the previously published methods. The proposed model is expected to significantly improve the accuracy of CBIR, contributing to enhanced image retrieval performance.

3. PROPOSED METHOD

In this section, we introduce a novel method that integrates multiple deep learning models to enhance image retrieval efficiency. Instead of relying on a single model, our approach combines results from multiple models and applies an adaptive weighting strategy to refine the final output. The proposed method is illustrated in Fig. 1.

In general, suppose there are n deep learning models, denoted as $M_1, M_2, \ldots, M_n$. Our ensemble method consists of the following steps.

Step 1. Evaluate the reliability of each model M_i

The reliability of a model reflects its accuracy in finding relevant images. Each deep learning model can have different accuracy depending on the dataset and image characteristics. Therefore,

using a reliability coefficient helps adjust the influence of each model during the result combination process. Models with high accuracy will have a greater influence in the image ranking process, helping to reduce the impact of poorly performing models. Additionally, if a model performs better on a specific type of image, it will be prioritized more, making the system more flexible.

The reliability of model M_i, denoted by C_i, is determined based on the accuracy of that model in image retrieval for the validation dataset. To fairly assess the contribution of each model within the ensemble, these values are normalized using the following formula

$$\hat{C}_i = \frac{C_i}{\sum_{j=1}^{m} C_j}, \tag{1}$$

where m is the number of models included in the ensemble. This normalization transforms the absolute accuracy values into relative weights that sum to 1, effectively reflecting the proportional contribution of each model. The normalized reliability scores $\hat{C}_i$ are subsequently used as weighting coefficients in the fusion strategy, enabling a balanced and performance-driven combination of model outputs.

Step 2. Construct the candidate image set

Each model M_i returns a list of retrieved images K_i, ordered by how similar they are to the query image Q. Our combination method creates a consolidated result set K by integrating the result sets K_i from each model.

$$K = \bigcup_{i=1}^{n} K_i. \tag{2}$$

Then, this set K is referred to as the candidate image set.

Step 3. Compute weights for images in the candidate set

To ensure the effectiveness of the set K, each image in this set will be assigned a weight to reflect the model's reliability, similarity to the query image, and the distributional characteristics of the images retrieved by the models.

For each $A \in K$, without loss of generality, we can assume that A is recommended by m models, where $m \leq n$. Then, the weight of the image A is calculated as follows

$$W(A) = \sum_{i=1}^{m} \hat{C}_i * S_i(A, Q) * \frac{1}{Std(A, K_i)}, \tag{3}$$

where $\hat{C}_i$ is the normalized reliability coefficient of the i^{th} model, computed as in Equation (1). A higher C_i, indicating better retrieval performance, leads to a larger $\hat{C}_i$, thereby increasing that model's contribution to the final weight $W(A)$; $S_i(A, Q)$ is a metric that measures the degree of similarity between image A and the query image Q; $Std(A, K_i)$ represents the standard deviation of distances from A to the images in K_i. This value is calculated as follows

$$Std(A, K_i) = \sqrt{\frac{1}{|K_i|} \sum_{B \in K_i} (d(A, B) - \mu)^2}, \tag{4}$$

where $d(A, B)$ is the distance between image A and each image B in the set K_i, μ is the average value of all distances from A to the images in the set K_i.

Step 4. Rank and return the results

After calculating the composite weight $W(A)$ for all images in the set K, the images are sorted in descending order of $W(A)$. The top k images with the highest weights are returned as the final retrieval results. In this way, our method not only enhances accuracy but also ensures the stability and performance of the system.

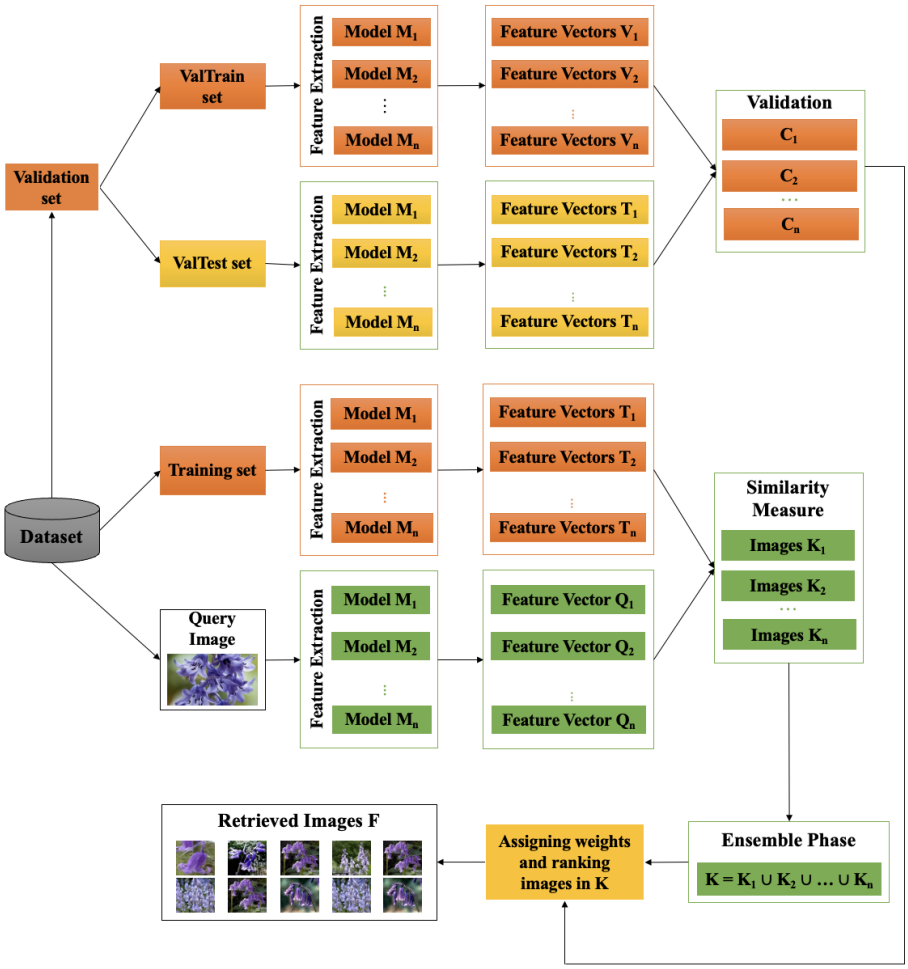


Figure 1: Proposed model diagram

4. EXPERIMENTS

4.1. Experimental setup

VgReEf-CBIR, which integrates VGG16, ResNet50, and EfficientNetB0. The purpose of this configuration is to test the suggested approach using a variety of traditional CNN designs. It offers a starting point for evaluating the weighted ensemble's performance with commonly used convolutional models in CBIR.

ReEfDe-CBIR, which integrates ResNet50, EfficientNetB0, and DenseNet201. This configuration replaces VGG16 with DenseNet201 to see if combining deeper and more current CNN architectures improves retrieval performance within the same ensemble framework.

SwCl-CBIR, which integrates Swin and Clip. This configuration combines two advanced architectures: Swin Transformer, which represents hierarchical vision transformers, and CLIP, a robust multimodal model. It is used to determine the scalability and applicability of the suggested strategy to cutting-edge deep representations.

These models leverage deep learning architectures pretrained on ImageNet to extract discriminative image features. Their effectiveness is analyzed by comparing retrieval performance against individual models, demonstrating the benefits of our proposed ensemble strategy.

We selected some benchmark datasets commonly used in CBIR, including Oxford-17-Flowers, Caltech-101, CIFAR-10, and ISIC-2018, as summarized in Table 1 and illustrated in Figure 2. Each dataset was split into training, validation, and testing subsets in a 70:20:10 ratio. Additionally, the validation set was further divided into ValTrain and ValTest in a 70:30 ratio for model evaluation, as described in Step 1 of the proposed method.

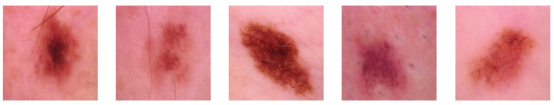
Table 1: Describe experimental datasets

Dataset Name	Images	Classes	Average Images per Class	Image Size
Oxford-17-Flowers	1,360	17	80	≈ 500 x 500 pixels
Caltech-101	9,144	101	40 - 800	Variable (not fixed)
CIFAR-10	10,000	10	1,000	32 x 32 pixels
ISIC-2018	10,015	7	115 - 6,705	600 x 450

(a) Oxford-17-Flower

(b) Caltech-101

(c) CIFAR-10



(d) ISIC-2018

Figure 2: Illustration of experimental datasets

All experiments were conducted using Python 3.9. The implementation was performed on a PC with an Intel(R) Core i7-11850H processor, 32GB RAM, a 1TB SSD, an RTX 3070 GPU, and Windows 11 Pro. For large datasets, the feature extraction process was carried out on Google Colab Pro using the T4 version. The system was developed using commonly adopted libraries in the Python ecosystem, including Scikit-learn, NumPy, OpenCV, Matplotlib, Pandas, and Transformers. For model execution, the TensorFlow framework was employed, especially for handling pretrained models such as CNNs and Vision Transformers. All input images were resized and normalized to 224×224 pixels for CNN-based models and 256×256 pixels for transformer-based models.

In CBIR, establishing effective evaluation metrics is essential for assessing both similarity between images and the overall retrieval performance. To quantify image similarity, cosine similarity is employed, as it measures the angle between feature vectors and ignores differences in their magnitudes. Meanwhile, retrieval effectiveness is evaluated using precision, which assesses the proportion of relevant images among the retrieved results. The formal definitions of these measures are presented below.

Given two vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ and θ the smallest angle between $\mathbf{a}$ and $\mathbf{b}$, the *cosine similarity* between $\mathbf{a}$ and $\mathbf{b}$, denoted $S(\mathbf{a}, \mathbf{b})$, is determined by

$$S(\mathbf{a}, \mathbf{b}) = \cos\theta = \frac{\mathbf{a}.\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}, \tag{5}$$

where $\mathbf{a}.\mathbf{b}$ is the dot product of vectors $\mathbf{a}$ and $\mathbf{b}$, calculated as Eq.(6), and $\|\mathbf{a}\|\|\mathbf{b}\|$ represents the product of the magnitudes (or Euclidean norms) of the vectors, calculated as Eq.(7).

$$\mathbf{a}.\mathbf{b} = \sum_{i=1}^{n} a_i b_i. \tag{6}$$

$$\|\mathbf{a}\| = \sqrt{(a_1^2 + a_2^2 + ... + a_n^2)}, \qquad \|\mathbf{b}\| = \sqrt{(b_1^2 + b_2^2 + ... + b_n^2)}. \tag{7}$$

Because of the smallest angle $\theta \in [0°, 180°]$, the value of cosine similarity ranges from -1 to 1.

Next, let Q_k denote the k-th query image.

The *precision* for Q_k, denoted $P(Q_k)$, is defined as

$$P(Q_k) = \frac{TP_k}{TP_k + FP_k}, \tag{8}$$

where TP_k represents the number of true positive predictions for Q_k, and FP_k denotes the number of false positive predictions for Q_k.

The *Recall* for the k-th query image Q_k, denoted as $R(Q_k)$, is defined as

$$R(Q_k) = \frac{TP_k}{TP_k + FN_k}, \tag{9}$$

where FN_k is the number of false negative predictions for query Q_k.

The *F1-score* for Q_k, denoted as $F1(Q_k)$, is the harmonic mean of Precision and Recall

$$F1(Q_k) = 2 \times \frac{P(Q_k) \times R(Q_k)}{P(Q_k) + R(Q_k)}. \tag{10}$$

To assess overall performance across the entire test set, we compute the average of these metrics over all N query images. The *mean precision (mP)*, *mean Recall (mR)* and *mean F1-score (mF1)* are respectively determined as

$$mP = \frac{1}{N}\sum_{k=1}^{N} P(Q_k), \quad mR = \frac{1}{N}\sum_{k=1}^{N} R(Q_k), \quad mF1 = \frac{1}{N}\sum_{k=1}^{N} F1(Q_k). \tag{11}$$

These aggregated metrics provide a more comprehensive evaluation of the system's ability to retrieve relevant content accurately and consistently.

4.2. Experimental results

In this section, we present the experimental results to evaluate the performance of the three proposed methods for combining deep learning models.

4.2.1. VgReEf-CBIR model

First, Table 3 shows the comparison results between the proposed VgReEf-CBIR model and individual models such as VGG16, ResNet50, and EfficientNetB0 on four datasets: Oxford-17-Flowers, Caltech-101, CIFAR-10 and ISIC-2018. It can be seen that the VgReEf-CBIR model achieves higher accuracy compared to individual models on most datasets, with 72.36% on the Oxford-17-Flowers dataset, significantly higher than the individual models. Specifically, VGG16 only achieved 63.09%, while ResNet50 and EfficientNetB0 had the same accuracy of 69.27%. Thus, the proposed method improves by 9.27% compared to VGG16 and 3.09% compared to both ResNet50 and EfficientNetB0.

In a similar way on the Caltech-101 dataset, VgReEf-CBIR continues to demonstrate superiority with an accuracy of 85.76%, compared to 80.64% of VGG16, 83.37% of ResNet50, and 84.77% of EfficientNetB0. Finally, on the CIFAR-10 dataset, VgReEf-CBIR achieved an accuracy of 77.93%, outperforming the top-performing individual model, ResNet50 (75.93%). This 2.0% improvement confirms the effectiveness of integrating models to improve the accuracy of the image retrieval system.

Table 2: Illustration on 10 random images from the Oxford-17-Flowers dataset

	VGG16			ResNet50			EfficientNetB0			VgReEf-CBIR		
Test mages	*Top10*	*Top20*	*Top30*	*Top10*	*Top20*	*Top30*	*Top10*	*Top20*	*Top30*	*Top10*	*Top20*	*Top30*
image_0093.jpg	80.00	45.00	30.00	80.00	50.00	36.67	60.00	45.00	36.67	**90.00**	**50.00**	**40.00**
image_0870.jpg	100.00	**100.00**	**86.67**	100.00	80.00	66.67	90.00	80.00	66.67	**100.00**	95.00	76.67
image_0488.jpg	90.00	65.00	50.00	100.00	95.00	76.67	100.00	**100.00**	**86.67**	**100.00**	95.00	83.33
image_0196.jpg	80.00	85.00	66.67	100.00	90.00	83.33	90.00	95.00	80.00	**100.00**	**95.00**	**90.00**
image_0515.jpg	70.00	65.00	56.67	70.00	60.00	50.00	70.00	50.00	43.33	**80.00**	**60.00**	**56.67**
image_0386.jpg	40.00	40.00	36.67	40.00	35.00	30.00	50.00	**45.00**	**40.00**	**70.00**	40.00	30.00
image_0132.jpg	20.00	30.00	26.67	70.00	50.00	**43.33**	60.00	35.00	33.33	**80.00**	**55.00**	40.00
image_0687.jpg	100.00	95.00	76.67	80.00	85.00	76.67	100.00	100.00	**100.00**	**100.00**	**100.00**	83.33
image_1056.jpg	70.00	55.00	**46.67**	50.00	35.00	33.33	40.00	50.00	36.67	**90.00**	**65.00**	43.33
image_0871.jpg	90.00	**90.00**	**93.33**	60.00	65.00	60.00	**100.00**	85.00	80.00	90.00	85.00	86.67

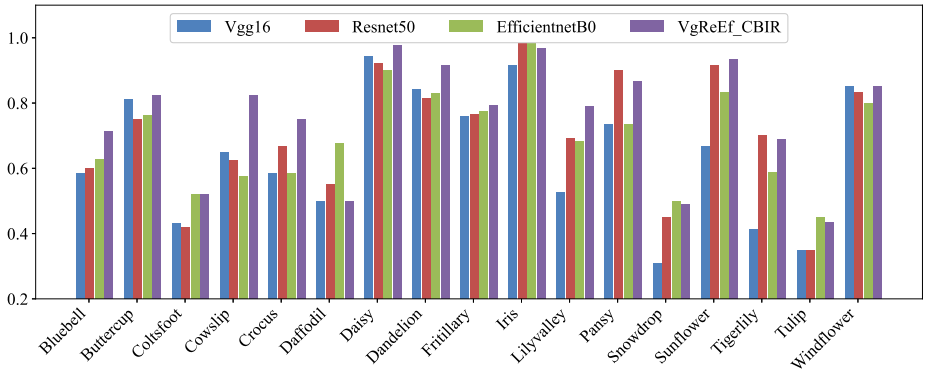


Figure 3: Compare the performance of the proposed model by labels on Oxford-17-Flowers

Furthermore, Figure 3 clearly shows that the VgReEf-CBIR model outperforms the other models in the Oxford-17-Flowers dataset for numerous kinds of labels. Similarly, Figure 4 shows that the proposed method continues to outperform the proposed model in accurately retrieving images based on labels, especially as the number of images in the CIFAR-10 dataset increases.

In addition, Table 2 shows the retrieval results for ten random images from the Oxford-17-Flowers dataset. The results indicate that the proposed method is more accurate when searching for similar

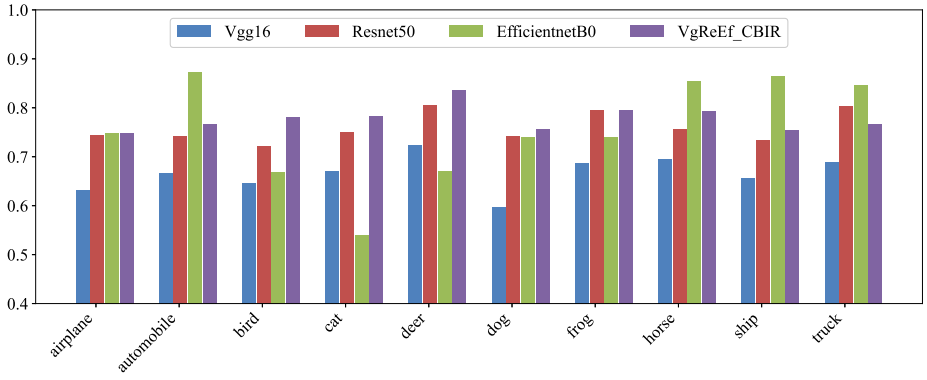


Figure 4: Compare the performance of the proposed model by labels on CIFAR-10

images, particularly when consistency criteria are considered in the search results. This allows the system to improve its reliability in practical applications.

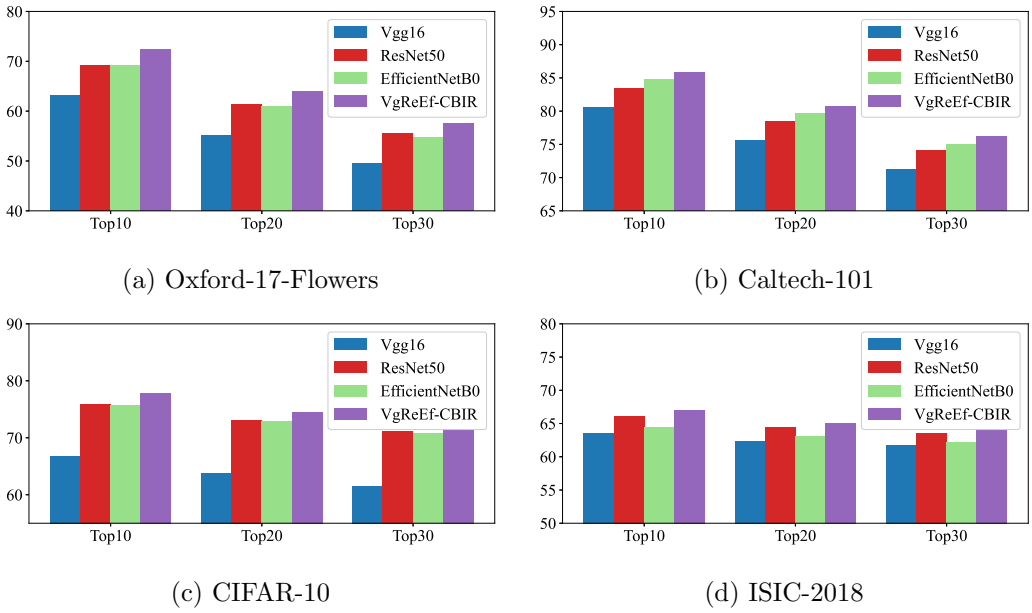
Experimental results show that the proposed VgReEf-CBIR model consistently outperforms individual models in terms of Recall and F1-score on most datasets and TopK. Specifically, using the Oxford-17-Flowers dataset at the Top30, the model earned a Recall of 95.16% and an F1-score of 68.75%, outperforming each individual component model, which had F1-scores ranging from 66.10 to 66.76%. Similarly, on Caltech-101, the F1-score at Top30 of VgReEf-CBIR reached 73.07%, up 0.76% to 4.07% from individual models. Even with hard datasets such as CIFAR-10 and ISIC-2018, the suggested model consistently improved F1-score at all TopK. These findings demonstrate that combining data from various models using a weighted combination process successfully leverages mutually supplemental features, considerably improving picture retrieval quality.

Finally, Figure 5 summarizes the performance of VgReEf-CBIR on the datasets. Observing the charts, we see that the proposed model shows significant improvement, especially when considering the Top10, Top20, and Top30. This further reinforces the assessment of the effectiveness of the feature combination method in the proposed model.

To summarize, on most datasets, the VgReEf-CBIR consistently outperformed single models. This result demonstrates the ensemble model's effectiveness in the image retrieval task while also pointing to new research directions that can be combined to improve retrieval accuracy.

Table 3: Results of the proposed VgReEf-CBIR model on the datasets

Dataset name	TopK	VGG16			ResNet50			EfficientNetB0			VgReEf-CBIR		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Oxford-17-Flowers	Top10	63.09	74.99	67.44	69.27	78.90	72.84	69.27	80.29	73.48	**72.36**	**81.74**	**75.87**
	Top20	55.12	83.58	64.37	61.38	89.57	70.40	60.89	90.16	70.44	**63.90**	**91.18**	**72.77**
	Top30	49.51	87.96	60.49	55.58	94.35	66.76	54.77	92.86	66.10	**57.64**	**95.16**	**68.75**
Caltech-101	Top10	80.64	60.48	60.87	83.37	62.70	63.35	84.77	63.92	64.84	**85.76**	**64.04**	**65.37**
	Top20	75.67	73.09	67.09	78.51	74.29	69.39	79.64	74.49	70.68	**80.76**	**74.60**	**71.24**
	Top30	71.29	79.87	69.00	74.17	80.85	71.46	74.99	80.60	72.31	**76.17**	**80.90**	**73.07**
CIFAR-10	Top10	66.81	7.45	13.40	75.93	8.49	15.27	75.76	8.48	15.24	**77.93**	**8.72**	**15.67**
	Top20	63.74	14.20	23.20	73.17	16.36	26.71	72.89	16.29	26.59	**74.50**	**16.65**	**27.18**
	Top30	61.49	20.53	30.73	71.13	23.85	35.66	70.86	23.74	35.50	**71.37**	**23.91**	**35.75**
ISIC-2018	Top10	63.55	2.47	4.23	66.09	2.67	4.59	64.49	2.46	4.23	**66.19**	**2.71**	**4.62**
	Top20	62.32	4.50	7.15	64.50	4.81	7.72	63.06	4.46	7.16	**64.51**	**4.97**	**7.83**
	Top30	61.77	6.40	9.69	63.63	6.87	10.39	62.20	6.27	9.60	**63.65**	**7.02**	**10.45**

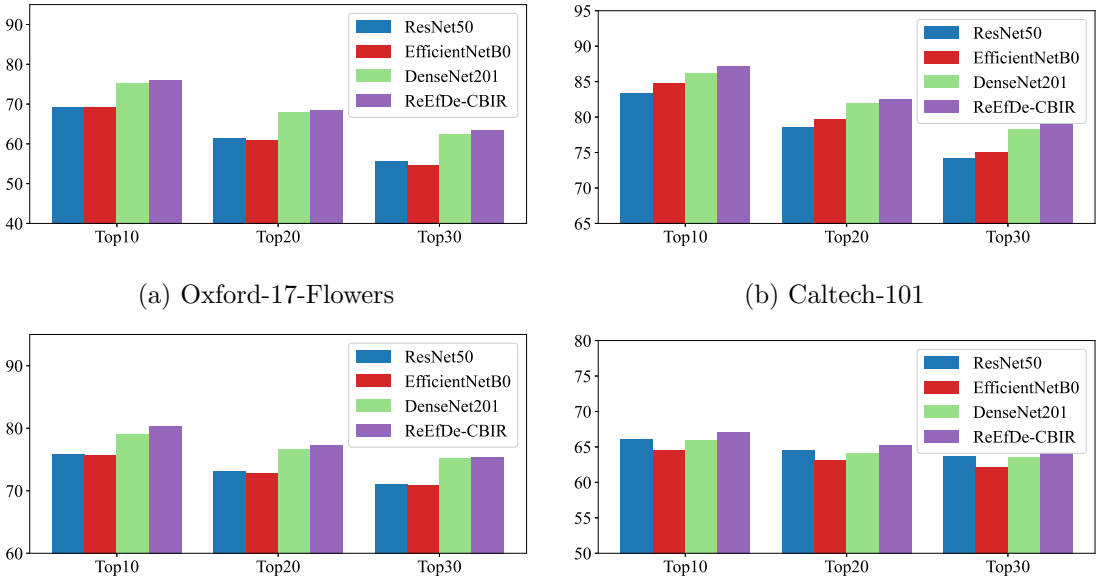


(a) Oxford-17-Flowers (b) Caltech-101

(c) CIFAR-10 (d) ISIC-2018

Figure 5: Illustration of the VgReEf-CBIR model performance on the datasets

4.2.2. ReEfDe-CBIR model



(a) Oxford-17-Flowers (b) Caltech-101

(c) CIFAR-10 (d) ISIC-2018

Figure 6: Illustration of the ReEfDe-CBIR model performance on the datasets

Table 4 shows the results of comparing the proposed ReEfDe-CBIR model with individual models such as ResNet50, EfficientNetB0, and DenseNet201. The results show that the proposed method

performs better, with significant improvements across all four datasets. On the Oxford-17-Flowers dataset, the ReEfDe-CBIR model achieved 75.53% at the Top10, outperforming ResNet50 and EfficientNetB0. Although DenseNet201 achieved 75.20%, the proposed method showed some improvement, confirming the effectiveness of feature combination in the CBIR model. Similarly, on the Caltech-101 dataset, the ReEfDe-CBIR model retained its advantage with an accuracy of 87.17%, outperforming ResNet50 (83.37%), EfficientNetB0 (84.77%), and DenseNet201 (86.13%). On the CIFAR-10 dataset, ReEfDe-CBIR outperformed DenseNet201 by 1.31%, achieving an accuracy of 80.41%. This improvement reinforces the proposed method's effectiveness when applied to a variety of datasets. Lastly, the ISIC-2018 dataset performs slightly better, which is understandable given its difficulty and complexity.

The proposed ReEfDe-CBIR model routinely beats the component models in both recall and F1-score across a variety of datasets, according to experimental data. The model's F1-score stayed strong, reaching 79.09% at Top10, while its Recall peaked at 97.33% at Top30 on the Oxford-17-Flowers dataset. A recall score of 81.40% and an F1-score of 75.38% at Top30 on Caltech-101 demonstrate the improvement in the coverage of relevant images, suggesting that the model not only recovers information reliably but also completely. With CIFAR-10, the suggested model beat single models despite of its high difficulty and low global recall, particularly in F1-score, a metric that is sensitive to the disparity between precision and recall, which achieved 28.24% at Top20 and 37.80% at Top30. The model's potential to handle increasingly complicated tasks is demonstrated by ReEfDe-CBIR's stability with the best metrics in the group on the ISIC-2018 dataset, despite its low absolute accuracy. These findings support the feature combination strategy's substantial advantages in raising retrieval efficiency overall.

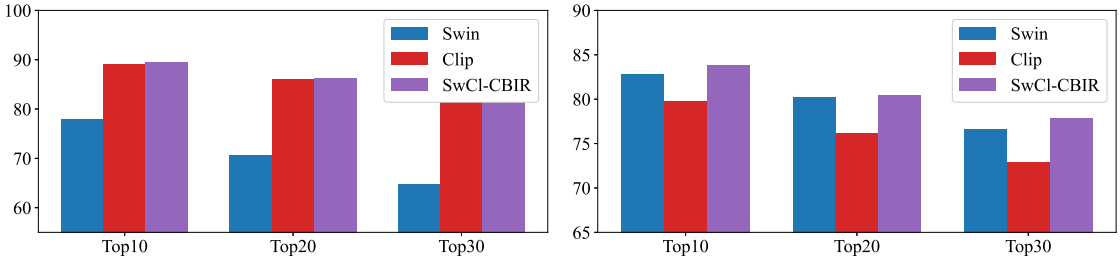
Furthermore, Figure 6 illustrates the performance of ReEfDe-CBIR, demonstrating the method's efficiency over individual models. Notably, the proposed model performed better on both the Caltech-101 and CIFAR-10 datasets, demonstrating greater stability and generalization capability than the other methods.

Table 4: Results of the proposed ReEfDe-CBIR model on the datasets

Dataset name	TopK	ResNet50			EfficientNetB0			DenseNet201			ReEfDe-CBIR		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Oxford-17-Flowers	Top10	69.27	78.90	72.84	69.27	80.29	73.48	75.20	83.31	78.29	**75.53**	**85.03**	**79.09**
	Top20	61.38	89.57	70.40	60.89	90.16	70.44	67.89	94.71	76.83	**68.50**	**94.79**	**77.10**
	Top30	55.58	94.35	66.76	54.77	92.86	66.10	62.44	97.33	73.38	**63.44**	**97.33**	**74.02**
Caltech-101	Top10	83.37	62.70	63.35	84.77	63.92	64.84	86.13	64.37	66.11	**87.17**	**64.96**	**66.70**
	Top20	78.51	74.29	69.39	79.64	74.49	70.68	82.00	75.03	72.44	**82.58**	**75.45**	**72.88**
	Top30	74.17	80.85	71.46	74.99	80.60	72.31	78.24	80.88	74.72	**79.00**	**81.40**	**75.38**
CIFAR-10	Top10	75.93	8.49	15.27	75.76	8.48	15.24	79.10	8.85	15.91	**80.41**	**9.00**	**16.17**
	Top20	73.17	16.36	26.71	72.89	16.29	26.59	76.70	17.15	27.99	**77.26**	**17.28**	**28.24**
	Top30	71.13	23.85	35.66	70.86	23.74	35.50	75.17	25.20	37.69	**75.39**	**25.28**	**37.80**
ISIC-2018	Top10	66.09	2.67	4.59	64.49	2.46	4.23	65.90	2.62	4.50	**66.80**	**2.78**	**4.74**
	Top20	64.50	4.81	7.72	63.06	4.46	7.16	64.11	4.75	7.53	**65.18**	**5.09**	**7.96**
	Top30	63.63	6.87	10.39	62.20	6.27	9.60	63.53	6.86	10.22	**64.20**	**7.19**	**10.63**

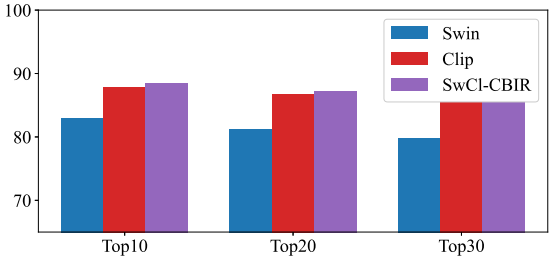
4.2.3. SwCl-CBIR model

The experimental results summarized in the Table 5 reveal that the proposed SwCl-CBIR method significantly outperformed the individual Swin and Clip models across all three evaluation metrics, Precision, Recall, and F1-score, on four widely used benchmark datasets.

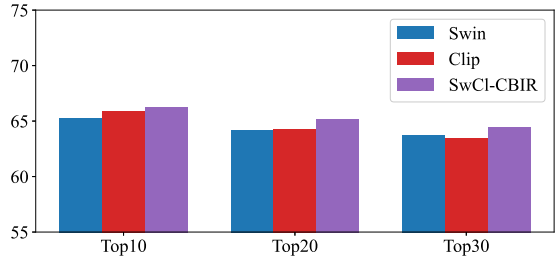


(a) Oxford-17-Flowers

(b) Caltech-101



(c) CIFAR-10



(d) ISIC-2018

Figure 7: Illustration of the SwCl-CBIR model performance on the datasets

On the Oxford-17-Flowers dataset, SwCl-CBIR achieved an F1-score of 91.51% at Top-10, which was notably higher than Swin (81.07%) and marginally better than Clip (91.01%). The proposed method also maintained a consistently strong performance across broader retrieval scopes, reaching 88.58% at Top-30, the highest among all evaluated models.

Similarly, on the Caltech-101 dataset, SwCl-CBIR yielded an F1-score of 74.14% at Top-30, reflecting an improvement of approximately 1.2% to 4.1% compared to the baseline Swin and Clip models.

In the case of CIFAR-10, which is known for its fine-grained categories and low recall rates, SwCl-CBIR still demonstrated measurable gains. At Top-30, the F1-score increased from 43.19% (Clip) to 43.25%, marking the highest score across all topK settings on this dataset.

On the ISIC-2018 medical imaging dataset, which presents substantial noise and domain complexity, the proposed model preserved its robustness. SwCl-CBIR improved the F1-score from 9.81% (Clip) to 10.34% at Top-30, indicating better resilience under challenging conditions.

These results confirmed that the integration of Swin and Clip features via a smart fusion mechanism in SwCl-CBIR effectively exploited the complementary strengths of both models. This synergy led to simultaneous improvements in retrieval accuracy, completeness, and balance between precision and recall. Consequently, SwCl-CBIR exhibits strong potential for real-world applications in CBIR tasks.

5. CONCLUSION AND FUTURE WORKS

To improve CBIR performance, this paper introduced an advanced ensemble method that synthesizes retrieval results using feature sets extracted from multiple deep learning models. The proposed ensemble process carefully evaluates the weight assigned to each retrieved image, considering not only the accuracy of the models on the validation dataset and image-query similarity, but also the

Table 5: Results of the proposed SwCl-CBIR model on the datasets

Dataset name	TopK	Swin			Clip			SwCl-CBIR		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Oxford-17-Flowers	Top10	78.05	85.98	81.07	89.02	94.16	91.01	**89.51**	**94.65**	**91.51**
	Top20	70.65	95.32	78.91	86.06	98.95	90.92	**86.22**	**98.95**	**91.11**
	Top30	64.69	96.94	74.67	81.73	99.44	88.39	**82.01**	**99.44**	**88.58**
Caltech-101	Top10	82.83	62.70	63.83	79.82	61.83	61.97	**83.79**	**63.58**	**64.96**
	Top20	80.19	73.24	70.76	76.21	72.45	68.04	**80.48**	**73.43**	**71.30**
	Top30	76.65	78.67	72.94	72.96	78.06	70.08	**77.86**	**79.48**	**74.14**
CIFAR-10	Top10	82.99	9.28	16.68	87.77	9.82	17.64	**88.49**	**9.90**	**17.79**
	Top20	81.18	18.16	29.64	86.71	19.39	31.65	**87.19**	**19.51**	**31.84**
	Top30	79.86	26.78	40.04	86.13	28.88	43.19	**86.26**	**28.93**	**43.25**
ISIC-2018	Top10	65.30	2.50	4.28	65.86	2.49	4.33	**66.22**	**2.63**	**4.47**
	Top20	64.19	4.56	7.24	64.28	4.52	7.32	**65.18**	**4.87**	**7.67**
	Top30	63.76	6.53	9.86	63.43	6.42	9.81	**64.42**	**6.93**	**10.34**

distributional characteristics of the images retrieved by models, as reflected by standard deviation. Building on this approach, we conducted experiments using several deep learning models, specifically VGG16, ResNet50, EfficientNetB0, DenseNet201, Swin, and Clip. These models, pre-trained on the ImageNet dataset, were repurposed for feature extraction on the Oxford-17-Flowers, Caltech-101, CIFAR-10, and ISIC-2018 datasets. We experimented with three model combinations within our proposed ensemble CBIR method framework: (1) VGG16, ResNet50 and EfficientNetB0 (VgReEf-CBIR), and (2) ResNet50, EfficientNetB0 and DenseNet201 (ReEfDe-CBIR), and (3) Swin and Clip (SwCl-CBIR). The results showed that our proposed method achieved superior image retrieval performance, outperforming the best individual model by 1% to 3% across most datasets. However, it is important to emphasize that selecting complementary deep learning models plays a crucial role in optimizing outcomes. We recommend choosing models with comparable accuracy levels to achieve more robust ensemble results.

In future studies, we plan to implement k-fold cross-validation to evaluate the model more comprehensively and reduce reliance on a single train-test split. Additionally, we will apply advanced statistical analysis methods, such as paired t-tests and confidence interval estimation for key performance metrics, to assess the stability and statistical significance of the results. Furthermore, we aim to enhance the proposed framework by integrating deep features from advanced architectures such as Clip, Swin Transformer, and DINO, while also exploring weighted ensemble strategies and multi-objective optimization techniques to further improve the effectiveness of the CBIR system.

ACKNOWLEDGMENT

The authors sincerely acknowledge support from the Faculty of Information Technology, University of Sciences, Hue University. Additionally, we extend our heartfelt gratitude to University of Khanh Hoa for providing the facilities essential for completing this research.

REFERENCES

[1] T. V. T. Le, T. T. Van, and T. M. Le, "An improvement of R-Tree for content-based image retrieval," in *Annals of the Eötvös Loránd University, Computer Science Section*, vol. 53, Budapest, Hungary, 2022. [Online]. Available: https://csdlkhoahoc.hueuni.edu.vn/data/2022/11/BB_Annales_Thanh.pdf

[2] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval: A review of recent trends," *Cogent Engineering*, vol. 8, no. 1, p. 1927469, 2021. [Online]. Available: https://doi.org/10.1080/23311916.2021.1927469

[3] Y. Mingqiang, K. Kidiyo, R. Joseph *et al.*, "A survey of shape feature extraction techniques," *Pattern Recognition*, vol. 15, no. 7, pp. 43–90, 2008.

[4] P. Enser and C. Sandom, "Towards a comprehensive survey of the semantic gap in visual image retrieval," in *International Conference on Image and Video Retrieval*. Springer, 2003, pp. 291–299. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-45113-7_29

[5] K. Ahmad, M. Sahu, M. Shrivastava, M. A. Rizvi, and V. Jain, "An efficient image retrieval tool: Query based image management system," *International Journal of Information Technology*, vol. 12, no. 1, pp. 103–111, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s41870-018-0198-9

[6] J. M. Patel and N. C. Gamit, "A review on feature extraction techniques in content based image retrieval," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2016, pp. 2259–2263.

[7] S. C. Ryszard, "Image feature extraction techniques and their applications for CBIR and biometrics systems," *International Journal of Biology and Biomedical Engineering*, vol. 1, no. 1, pp. 6–16, 2007.

[8] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian, "CNN vs. SIFT for image retrieval: Alternative or complementary?" in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 407–411. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2964284.2967252

[9] R. Kapoor, D. Sharma, and T. Gulati, "State of the art content based image retrieval techniques using deep learning: a survey," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 29 561–29 583, 2021.

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (Csur)*, vol. 40, no. 2, pp. 1–60, 2008.

[11] S. Kumar, M. K. Singh, and Mishra, "Improve content-based image retrieval using deep learning model," *Journal of Physics: Conference Series*, vol. 2327, no. 1, p. 012028, 2022.

[12] E. Ranjith, L. Parthiban, T. P. Latchoumi, S. A. Kumar, D. G. Perera, and S. Ramaswamy, "An effective content based image retrieval system using deep learning based inception model," *Wireless Personal Communications*, vol. 133, no. 2, pp. 811–829, 2023.

[13] G. Shamsipour, S. Fekri-Ershad, M. Sharifi, and A. Alaei, "Improve the efficiency of handcrafted features in image retrieval by adding selected feature generating layers of deep convolutional neural networks," *Signal, Image and Video Processing*, vol. 18, no. 3, pp. 2607–2620, 2024.

[14] A. Soni and A. Kumar, "Deep learning-based image retrieval: Addressing the semantic gap for accurate content-based image retrieval," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*. IEEE, 2023, pp. 1512–1517.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] P. Singh, P. Hrisheekesha, and V. K. Singh, "CBIR-CNN: Content-based image retrieval on celebrity data using deep convolution neural network," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 14, no. 1, pp. 257–272, 2021.

[17] K. Zhang, S. Qi, J. Cai, D. Zhao, T. Yu, Y. Yue, Y. Yao, and W. Qian, "Content-based image retrieval with a convolutional siamese neural network: Distinguishing lung cancer and tuberculosis in CT images," *Computers in Biology and Medicine*, vol. 140, p. 105096, 2022.

[18] A. Eswaran and E. Varshini, "Reverse image search engine for garment industry," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2022, pp. 414–418.

[19] M. R. Reena and P. Ameer, "A content-based image retrieval system for the diagnosis of lymphoma using blood micrographs: An incorporation of deep learning with a traditional learning approach," *Computers in Biology and Medicine*, vol. 145, p. 105463, 2022.

[20] A. Sharma, V. Singh, and P. Singh, "Deep CNN based hybrid model for image retrieval." [Online]. Available: https://www.ijitee.org/portfolio-item/g92030811922/

[21] V. Anand, S. Gupta, D. Gupta, Y. Gulzar, Q. Xin, S. Juneja, A. Shah, and A. Shaikh, "Weighted average ensemble deep learning model for stratification of brain tumor in MRI images," *Diagnostics*, vol. 13, no. 7, p. 1320, 2023.

[22] N. Mungoli, "Adaptive ensemble learning: Boosting model performance through intelligent feature fusion in deep neural networks," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.02653

[23] Z. Ullah and J. Kim, "Hierarchical deep feature fusion and ensemble learning for enhanced brain tumor MRI classification," *Mathematics*, vol. 13, no. 17, p. 2787, 2025.

[24] I. T. Chughtai, A. Naseer, M. Tamoor, S. Asif, M. Jabbar, and R. Shahid, "Content-based image retrieval via transfer learning," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 5, pp. 8193–8218, 2023.

Received on March 10, 2025
Accepted on August 13, 2025