International Journal of Basic and Applied Sciences, 14 (5) (2025) 87-97

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/scxj1q69
Research paper

An enhanced CBIR System Using Modified VGG16 with Stacked SVM, Random Forest, and XGBoost Classifiers

S. Ravi *, Kamal Sutaria

Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat, India
**Corresponding author E-mail: 2123004136006@paruluniversity.ac.in*

Received: July 15, 2025, Accepted: August 26, 2025, Published: September 4, 2025

Abstract

CBIR systems have improved large image dataset search and management, yet traditional techniques with HSV GLCM and SIFT experience limited precision because of their insufficient feature extraction abilities. The research requires additional efforts to boost retrieval precision and the overall computer vision model functioning, specifically for CNN-SVM constructions. Through VGG16-based feature extraction, a newly proposed hybrid Content-Based Image Retrieval framework incorporates a stacked classification system built using SVM alongside RF and XGBoost for decision-making. Tests on the Wang dataset validated this model by showing the CNN-SVM baseline model delivering precision levels of 83.61% at 10 retrievals, and both 83.67% at 15 and 83.37% at 20 retrievals. Implementing stacked classifiers produced advanced decision boundaries that improved classification performance while raising total precision by 12.06% at all retrieval settings above the baseline model. The research work establishes the basic groundwork for next-generation hybrid CBIR approaches while demonstrating the benefits of deep learning and ensemble technologies in improving retrieval performance.

Keywords: *Convolutional Neural Network (CNN); Content-Based Image Retrieval (CBIR); Extreme Gradient Boosting (XGBoost); Random Forest (RF).*

1. Introduction

The rapid expansion of digital content has created an increasing need for efficient techniques to search and retrieve images from large datasets. Content-based image retrieval systems manage digital content through visual features, including color, texture, and shape functions without depending on text-oriented labels [1]. Handcrafted retrieval features in former CBIR systems included Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) [2], and Gray Level Co-occurrence Matrix (GLCM), yet produced poor semantic recognition results [3]. Deep learning technology [4] created a breakthrough in CBIR systems because it allows computers to automatically harvest strong deep features encoded with meaning.

Convolutional Neural Networks (CNNs) [5] demonstrate outstanding capabilities for extracting complex hierarchical features, thus making them popular for CBIR applications.

The pre-trained VGG16 model shows excellence in feature extraction when used for image classification and retrieval since it was trained on the ImageNet database [6-7]. SoftMax classification prevents CNNs from performing fully optimized feature extraction because it shows limitations when processing complex decision boundaries. The research combined SVMs classifies using CNNs as an approach to achieve better accuracy levels. The CNN-SVM base model [8] demonstrated better performance than standard approaches, which employed HSV and GLCM features. SVM functions effectively as a single classifier, yet its performance constraints reduce its adaptability toward diverse classification tasks and complex image patterns [9]. The image retrieval system integrates VGG16 features along with SVM, Random Forest (RF) [10], and XGBoost [11] to work at different levels of classification, which helps improve accuracy and reliability. The combination of multiple classifiers improves retrieval effectiveness through linear and non-linear data correction that optimizes feature manipulation and classification mechanisms. A hybrid model containing stacked classifiers combined with VGG16 feature extraction methods has been developed to analyze the Wang dataset at retrieval thresholds of 10, 15, and 20. Results demonstrate major accuracy enhancements, which led to improved system performance while delivering better reliability at every retrieval depth.

2. Related work

Content-Based Image Retrieval (CBIR) research continues to evolve throughout many years because researchers must retrieve essential images from extensive databases by analyzing visual content [12-13]. The development of CBIR methods can be broadly grouped into three main categories: The field has developed into three main approach categories, including handcrafted feature-based techniques, deep learning-based models, and systems that bring together features from both methods [14].

2.1. Handcrafted feature-based methods

Prior computer-based image retrieval technologies depended mainly on color and texture patterns and geometric forms to construct image outlines. Technical approaches, including Histogram of Oriented Gradients (HOG) [15] and Scale-Invariant Feature Transform (SIFT) and Gray-Level Co-occurrence Matrix (GLCM) [16], showed remarkable results. The approach faced difficulties in establishing meaningful associations between fundamental image attributes and human concept recognition, leading to restricted retrieval capabilities across various image collections.

2.2. Deep learning-based approaches

The development of deep learning enabled automatic detection of vital image features during retrieval operations. The networks AlexNet, VGG16, ResNet [17], and Inception [18] establish powerful features during extraction processes. The ImageNet pre-trained VGG16 model serves as a ubiquitous choice for the implementation of transfer learning techniques in image retrieval systems [19]. Services based on machine learning models no longer require manual feature design since they present adaptive solutions throughout multiple domains [20]. Software Maximum classification, which powers CNNs, faces problems defining exact separation lines between classes.

2.3. Hybrid models

Users experience enhanced graphic pattern detection through the usage of a CNN-SVM hybrid system, which connects CNN feature extraction with SVM classification [21]. The concept developed into ensemble models [22], which combined Random Forests with Gradient Boosting Machines and CNN-based features across diverse data distributions [23]. The union of multiple computer vision techniques enhances both system operational accuracy and performance measurements in image retrieval platforms

2.4. Challenges in existing systems

While hybrid CBIR models have significantly advanced retrieval accuracy, several challenges remain:

- Relying solely on a single classification model (for example, Support Vector Machine) restricts the system's ability to accommodate a variety of complex data sets.
- Conventional ensemble methods frequently struggle to effectively manage inter-class connections, resulting in suboptimal retrieval for categories with unclear boundaries.
- Current models often fail to incorporate a multi-layered classification system that leverages the unique capabilities of various classifiers.

Implemented CBIR techniques have several limitations despite their advancement. Handcrafted descriptors such as GLCM and SIFT are useful in terms of low-level feature type, but do not capture semantic information, leading to poor performance on complex visual concepts [16]. The addition of the CNN in some instances enhanced the performance of the SVMs, but frequently the decision boundaries were non-linear, causing the hybrid to fail to classify in categories where the descriptive patterns overlapped [21]. This was addressed by the Ensemble framework [22-23], yet there was no hierarchical combination of classifiers, thereby limiting generalization. The proposed stacked classifier attempts to alleviate these limitations by nesting linear (SVM) with the non-linear (RF, XGBoost), decision boundaries, and capturing complementary decision boundaries, increasing precision. Recent works have ventured into attention-based CBIR [22, 25] and transformer-based retrieval models [26-27], a demonstration of the relevance of semantic weighting. It is in this direction that this present work integrates the approaches of stacked classifiers with attention-enhanced feature extraction to achieve both robustness and scale.

2.5. Motivation for proposed work

A combined approach using VGG16 feature extraction with stacked classifiers leads to better CBIR systems by improving retrieval performance. The system integrates multiple classifiers into one unified platform to solve research problems while boosting operational picture searching across all categories.

3. Proposed methodology

A new CBIR system emerges due to the combination of CNNs and machine learning classifiers in this research. The Wang dataset receives performance evaluation using defined metrics from the base model, which follows a specified methodological framework.

3.1. Dataset preparation

Although widely used as a benchmark, the Wang dataset has intra-class diversity that bears heavily on retrieval complexity. To take a specific example, within the category of Africa, the content is heterogeneous, so we find not only the landscape, but also wildlife and cultural motifs, providing a significant degree of variation within the same category. In the same way, Buildings contains historical and modern buildings and hence different representations of features. This diversity contributes to the effectiveness of evaluation, but the size of the dataset is rather small, and this aspect may limit the applicability of findings to bigger repositories. The potential limitations of the proposed hybrid CBIR system are the insufficient size of the datasets used as well as their limited diversification, which future work will address by testing on larger and more diverse datasets, e.g., ImageNet or Places365, thereby verifying the scalability and robustness of the proposed hybrid CBIR system.

The Wang dataset contains 1000 images that provide the standard benchmark for CBIR research by organizing the collection into 10 standardized classes, such as Africa, Beaches, Buildings and Buses, and Dinosaurs. Each 100-item category within the dataset receives pre-processing treatment that standardizes resolutions to 224×224 pixels before normalization of pixel values for base model requirements. Programming schemes add correct product descriptions to image collections.

3.2. Feature extraction with modified VGG16

The proposed model applies a modified VGG16 network trained on ImageNet by replacing the classification layers with a GAP [24] layer to extract features instead of performing identification. The setup obtains robust high-dimensional feature vectors alongside data compression. Next, the model continues training for domain adaptation on Corel images to acclimate to specific traits there.

3.3. Hybrid classifier design

The hybrid classification model combines the strengths of multiple classifiers for accurate image retrieval.

- SVM uses its strength in handling datasets of diverse dimensions to classify extracted content features for CBIR purposes.
- The ensemble methodology of Random Forest brings robust performance by efficiently addressing data vectors corrupted by noise.
- XGBoost processes information through the integration of SVM and Random Forest outputs, which possesses base classifier strengths to boost search results quality.

3.4. Similarity measurement and retrieval

Retrieval operations depend heavily on similarity ranking, which operates between both database images and query images, while primary similarity measures function as system standards.

- The similarity assessment between the database and query image vectors uses Euclidean Distance through direct line measurement of their combined features.

The top n images with the highest similarity scores are retrieved. For evaluation, results are obtained with n=10, n=15, and n=20.

3.5. Evaluation metrics

The General model is evaluated using the following metrics:

- Precision: The proportion of actual images in relation to total image retrieval results.
- Recall: Relevant image retrieval percentage computed against the entire universe of relevant dataset images.
- Accuracy: The overall proportion of correctly classified images, computed as

$\text{Accuracy}=(\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$.

The proposed model metric is on precision, calculated for each category of the Wang dataset and averaged for comparison with the base model.

3.6. Comparison with base model

Experimental results demonstrate that the hybrid CBIR system delivers better detection than its CNN-SVM base model, along with traditional methods, including HSV, GLCM, and SIFT. Experimental data reveal that accuracies rise dramatically when testing includes sample quantities between 10 to 20.

3.7. General structure of proposed model pipeline

3.7.1. Dataset details

Working with the Wang dataset, containing 1000 images, where images are labelled in the following manner:

- 0–99: Label 0
- 100–199: Label 1
- 200–299: Label 2, and so on up to Label 9.
- Images are stored in .jpg format.

3.7.2. Components of the system

- Feature Extraction:
- A modified VGG16 model is used to extract deep features from Images.
- The "include top" parameter gets set to False, which allows us to drop fully connected layers while extracting features from convolutional operations.
- Process features become flattened structures, which get saved for future training and retrieval operations.
- Classification:
- Base Classifiers:
- Support Vector Machine (SVM) with probability estimation enabled.
- Random Forest Classifier.
- Meta-Classifier:
- There is a use case for XGBoost in which it operates as a stacked classifier to produce final predictions from SVM and Random Forest base models.
- The meta-classifier gains features through the utilization of base classification model predictions.
- Image Retrieval:
- After feature extraction, for a given query image:
- Its features are compared to the pre-computed dataset features using the Euclidean distance.
- The top-n most similar images are retrieved and displayed.

The fig. 1, illustrates the overall architecture of the proposed CBIR system. It includes the stages of feature extraction using modified VGG16, classification through stacked classifiers (SVM, RF, XGBoost), and image retrieval based on similarity scores.

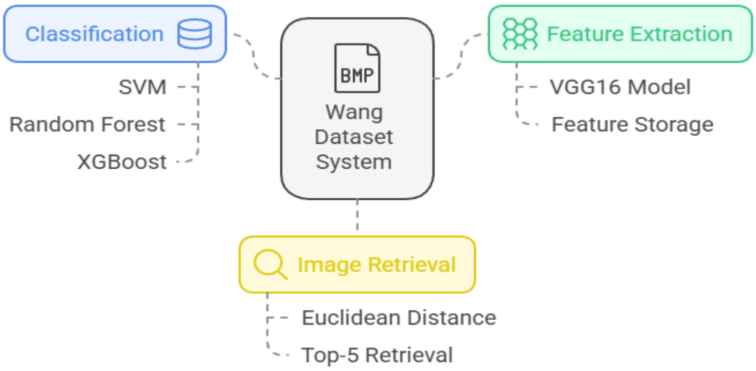


Fig. 1: Brief Proposed Model Pipeline.

3.8. Algorithm for the proposed hybrid CBIR model

Below is the algorithmic representation of the proposed methodology:

```
Algorithm: Hybrid CBIR Model
Input: Query image Q, Image dataset D, Number of retrieved images n.
Output: Top n similar images, Precision, Recall, and Accuracy.
1)  Pre-processing:
2)  1.1. Resize all images in D and Q to 224×224 pixels.
3)  1.2. Normalize the pixel values to a range of [0, 1].
4)  Feature Extraction:
5)  2.1. Load the pre-trained VGG16 model (with ImageNet weights).
6)  2.2. Remove the top classification layers and add a Global Average Pooling (GAP) layer.
7)  2.3. Extract feature vectors $F_Q$ for Q and $F_D$ for all images in D.
8)  Hybrid Classifier Training:
9)  3.1. Split D into training (60%) and testing (40%) subsets.
10) 3.2. Train a linear SVM on the feature vectors $F_{D.}$
11) 3.3. Train a Random Forest (RF) classifier on $F_{D.}$
12) 3.4. Combine predictions of SVM and RF to train the XGBoost meta-classifier.
13) Image Retrieval:
14) _# Compute the similarity of FQ with FD using:
15)  - Euclidean Distance.
16) # Rank all images in D based on similarity scores.
17) # Retrieve the top n images.
18) Evaluation:
19)  # Compare retrieved labels with the ground truth label of Q.
20) # Calculate metrics:
21)  - Precision
22)  - Recall
23)  - Accuracy
24) # Repeat steps 4–5 for all test query images.
25) Comparison with Baseline Models:
26) 6.1. Evaluate the performance of the proposed model and compare it with the base model and traditional CBIR methods.
27) Output:
28) # Display query image and retrieved top n images.
29) # Report the precision, recall, and accuracy for all categories.
```

3.9. Mathematical representation of the proposed system

3.9.1. Feature extraction using modified VGG16

Let the input image be denoted as:

$$I \in R^{224\times224\times3} \tag{1}$$

The image is passed through the modified VGG16 model with a Global Average Pooling (GAP) layer to extract the feature vector:

$$F= GAP\ (VGG16\ (I)) \in R^{1\times d} \tag{2}$$

Where:

- F represents the extracted feature vector
- d is the dimensionality of the vector

3.9.2. base classifier predictions

The extracted feature vector F is fed into two base classifiers:

- Support Vector Machine (SVM):

$h_{SVM}(F) \in R^k$ (3)

- Random Forest (RF):

$h_{RF}(F) \in R^k$ (4)

Where k is the number of output class probabilities or predictions.

3.9.3. Meta-classifier (XGboost)

XGBoost operates as the stacked meta-classifier, combining the outputs from SVM and RF:

$Z = [\, h_{SVM}(F), h_{RF}(F)] \in R^{2k}$ (5)

Final prediction:

$y = h_{XGB}(Z)$ (6)

Where:

- Z is the concatenated feature vector from base classifiers
- y is the final class prediction from XGBoost

3.9.4. Similarity measurement (Euclidean distance)

For retrieval, the similarity between a query image Fq and a dataset image F_d is measured using Euclidean distance:

$$D(F_q, F_d) = \sqrt{\sum_{i=1}^{d} \left(F_q^{(i)} - F_d^{(i)}\right)^2} \quad (7)$$

The top-n most similar images (lowest distances) are returned as retrieval results.
The fig. 2 illustrates the complete pipeline of the proposed hybrid CBIR system. It begins with pre-processing and feature extraction using a modified VGG16 network. The extracted features are passed through stacked classifiers (SVM, Random Forest, and XGBoost) to predict image labels. Finally, similarity is computed between query and dataset images using Euclidean distance to retrieve the top-matching images.

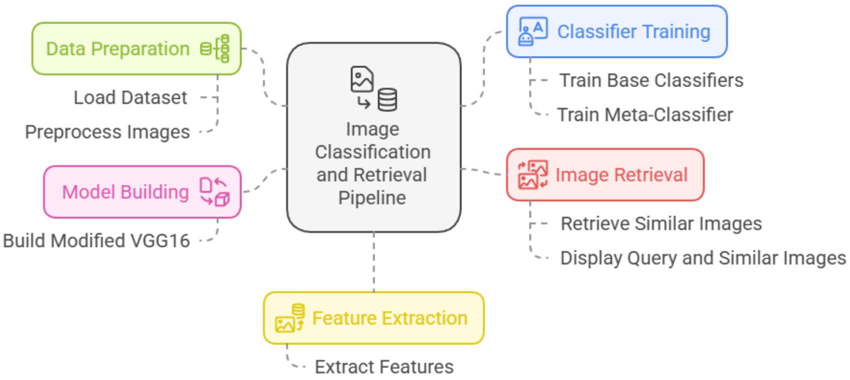


Fig. 2: Image Classification and Retrieval Pipeline.

4. Experimental results

Experimental results from the hybrid CBIR system utilize the Wang dataset, which comprises ten classes with 100 images each: Africa, beaches, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, and food.

4.1. Experimental setup

The system was constructed from Python code alongside image processing libraries, which integrated building blocks for feature extraction and classification functions. Research work generated the system according to the established technical requirements:

- Processor: Intel Core i7, 11th Generation
- RAM: 16 GB
- Operating System: Windows 10
- Programming Environment: Anaconda with Python 3.8

Regarding the efficiency of the computations, an hour and a half was needed to train the stacked classifier, the computing parameters: Intel Core i7 (11th Gen) eight-core processor, 16 GB of RAM; no graphics card. The time of inference by a single query image averaged 0.54 seconds, a time that is acceptable to medium-scale applications. In the real-time or mass implementation scenarios, however, it would need such optimization techniques as to use a GPU-based acceleration, pruning, or quantization. This shows the compromise between retrieval quality and practicality of computation in resource-scarce systems.

4.2. Evaluation metrics

The proposed model determines its efficiency by utilizing precision metrics. The performance metrics were evaluated across different image set retrieval numbers from 10 to 15 to 20 to determine how performance changes with increasing quantity.

4.3. Comparison with the base model

A comparison of the proposed model appears with data from the base paper that demonstrated CNN-SVM algorithm compatibility. The evaluation methodology displays precision assessment details in Table 4 for each category.

4.4. Results for different retrieval sizes

Table I summarizes the precision values for the proposed model when retrieving N = 10, 15, and 20 similar images. The results are then compared with the base paper's precision values for the same categories.

Table 1: Summarizes the Precision Values on Several Retrieval Images

Number of retrieved similar images(n)	Base Paper Precision (%)	Proposed Model Precision (%)
N=10	83.61	98.00
N=15	83.67	95.33
N=20	83.37	93.5

In Fig 3, the bar graph visually compares the precision scores of the base and proposed models for varying retrieval sizes. It emphasizes the superior performance of the proposed hybrid CBIR system.

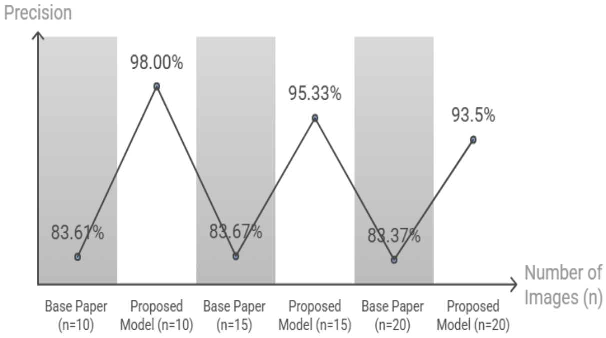


Fig. 3: Bar Graph Comparing the Accuracy Rate of the Proposed Hybrid CBIR Model with the Baseline CNN-SVM at Retrieval Sizes of N = 10, 15, and 20. The Hybrid Model Demonstrates A Higher Level of Precision with Evident Gains in All the Retrieval Depths.

From Table 1 and Fig. 3, it is evident that the proposed model significantly outperforms the base model for all retrieval sizes.

4.5. Per-class results

Table II shows the precision for each category in the Wang dataset at retrieval sizes of 10, 15, and 20. It provides a detailed view of category-wise performance and highlights high-performing classes like Dinosaurs, Flowers, and Buses.

Table 2: Precision Metrics of Categories Using the Wang Dataset

Category with Class number	The precision obtained relative to the number of retrieved similar images (in percentage (%)) for each category		
	Precision (n=10) in (%)	Precision (n=15) in (%)	Precision (n=20) in (%)
Africa (0)	90	66.67	60
Beaches (1)	100	100	100
Buildings (2)	90	93.33	95
Buses (3)	100	100	100
Dinosaurs (4)	100	100	100
Elephants (5)	100	100	100
Flowers (6)	100	100	100
Horses (7)	100	100	100
Mountains (8)	10	100	100
Food (9)	100%	93.33	80

In Fig. 4, the graph presents the top-performing categories in terms of precision for different retrieval sizes. It demonstrates how the proposed model maintains high accuracy in visually distinct classes.

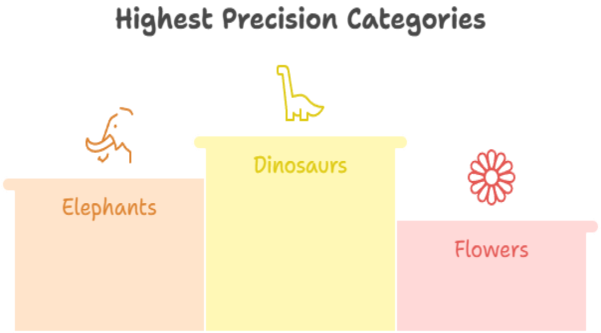


Fig. 4: Highest Precision Categories Based on the N Retrieval Values of 10, 15, and 20.

Table III compares the proposed model's precision with other feature extraction and classification techniques, such as HSV, GLCM, SIFT, CNN, and CNN-SVM. It confirms the superior performance of the stacked classifier system.

Table 3: Comparative Performance

Model	HSV	GLCM	SIFT	CNN	CNN-SVM	CNN with Stack Classifiers
Precision	0.43	0.39	0.53	0.79	0.84	0.95

In Fig. 5, the bar chart visually represents the precision scores from Table III. It clearly shows that the proposed CNN with stacked classifiers significantly outperforms traditional and CNN-based approaches.

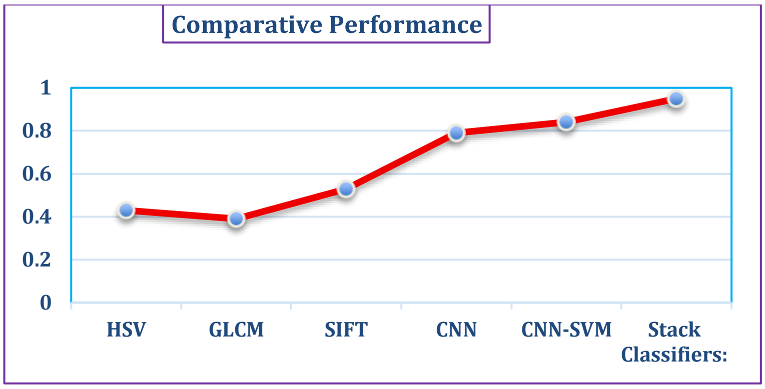


Fig. 5: Model Performance Comparison.

4.6. Visual results

A visual representation of the image retrieval outcomes can be seen in Fig. 6 with complementary images to assess system execution based on human perception. Retrieved images appear in a sorted list based on similarity scores created through a hybrid system that integrates feature extraction with classification methods. The fig. 6 presents query images and the system's retrieval of their top-10 matching images.

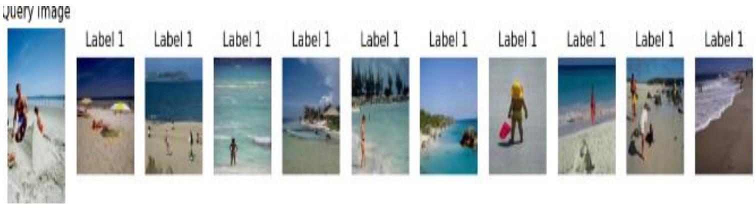
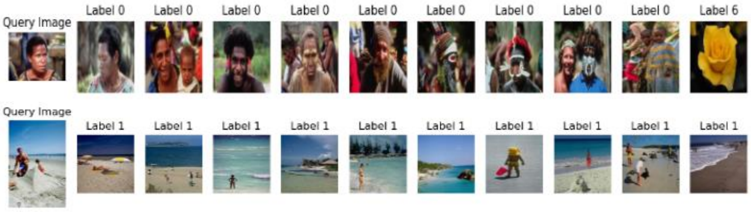


Fig. 6: Query Images with 10 Retrieval Images.

The Fig 7, 8, and 9 demonstrate retrieval results for increasing values of N. They showcase the effectiveness of the system in retrieving relevant images consistently as the retrieval size increases.

Figure 7 presents the image retrieval results when the system retrieves the top 10 similar images. The retrieved images exhibit strong visual relevance to the query image. This demonstrates the model's high precision in identifying top-ranked matches. It highlights the system's robustness at smaller retrieval sizes.



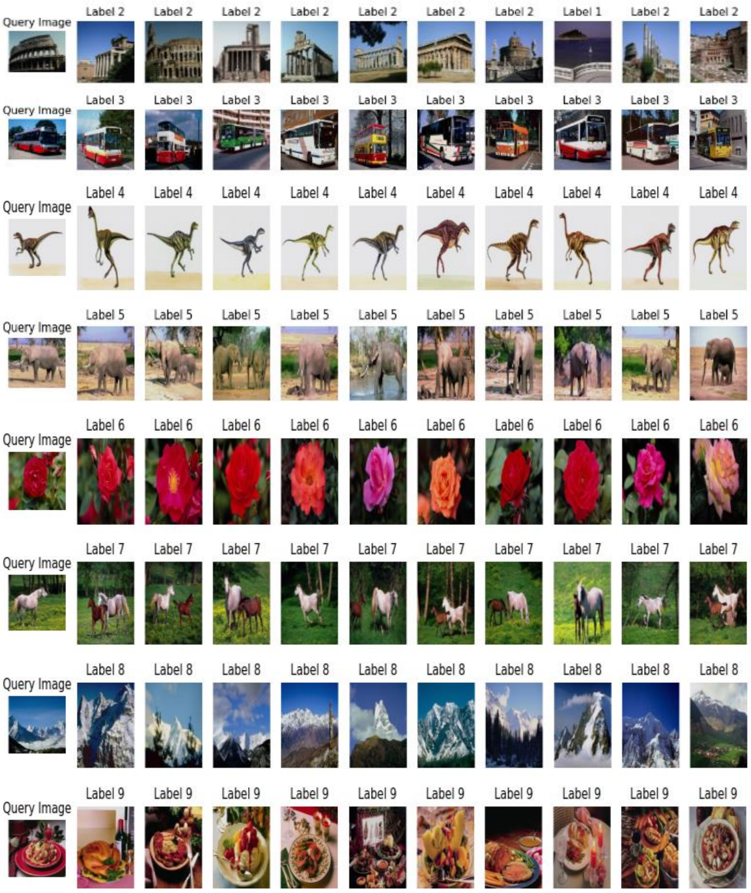
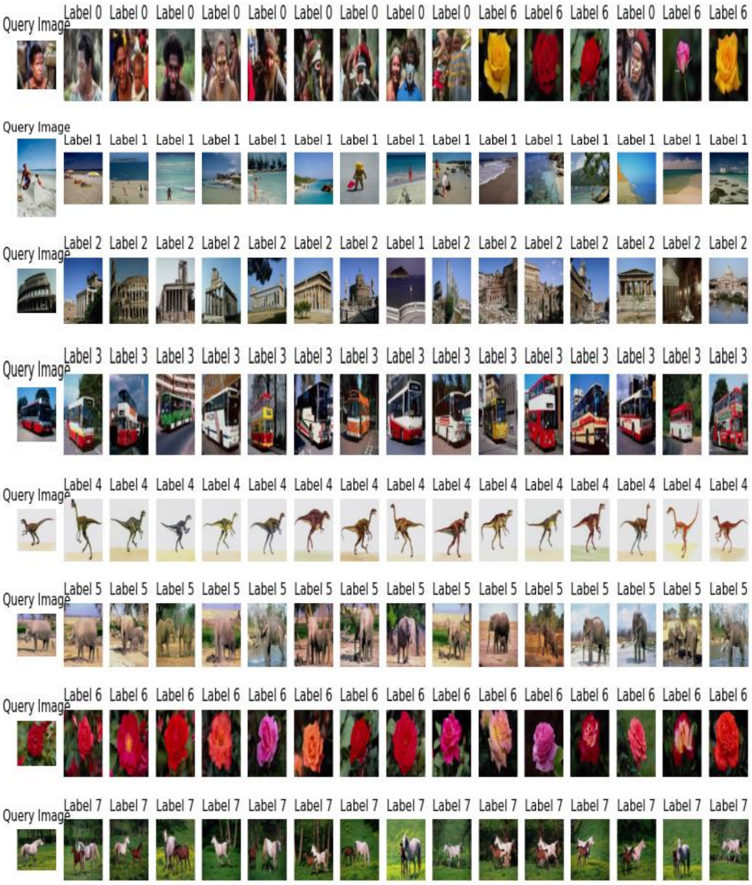

Query Image
Label 2 Label 2 Label 2 Label 2 Label 2 Label 2 Label 2 Label 1 Label 2 Label 2
Query Image
Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3
Query Image
Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4
Query Image
Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5
Query Image
Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6
Query Image
Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7
Query Image
Label 8 Label 8 Label 8 Label 8 Label 8 Label 8 Label 8 Label 8 Label 8 Label 8
Query Image
Label 9 Label 9 Label 9 Label 9 Label 9 Label 9 Label 9 Label 9 Label 9 Label 9

Fig. 7: Query Image Retrieval Results at N=10.

Figure 8 displays the retrieval performance for the top 15 similar images. The system continues to retrieve visually consistent and relevant images across most categories. It reflects the model's ability to maintain retrieval accuracy as the number of results increases. Precision remains high for visually distinct classes.


Query Image
Label 0 Label 0 Label 0 Label 0 Label 0 Label 0 Label 0 Label 0 Label 0 Label 6 Label 6 Label 6 Label 0 Label 6 Label 6
Query Image
Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1 Label 1
Query Image
Label 2 Label 2 Label 2 Label 2 Label 2 Label 2 Label 2 Label 1 Label 2 Label 2 Label 2 Label 2 Label 2 Label 2 Label 2
Query Image
Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3 Label 3
Query Image
Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4 Label 4
Query Image
Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5 Label 5
Query Image
Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6 Label 6
Query Image
Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7 Label 7

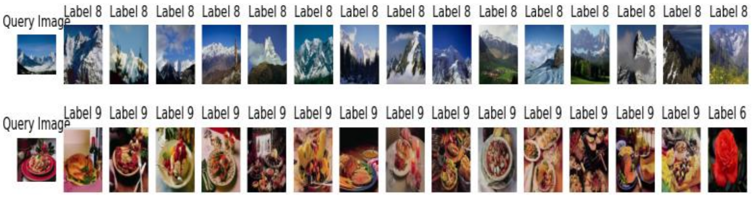


Fig. 8: Query Image Retrieval Results at N=15.

Figure 9 shows the retrieval output when 20 similar images are retrieved per query. Despite the increased retrieval size, the model sustains good relevance across most images. This confirms the scalability and generalization strength of the hybrid CBIR system. It effectively manages more diverse retrieval cases without a significant precision drop.

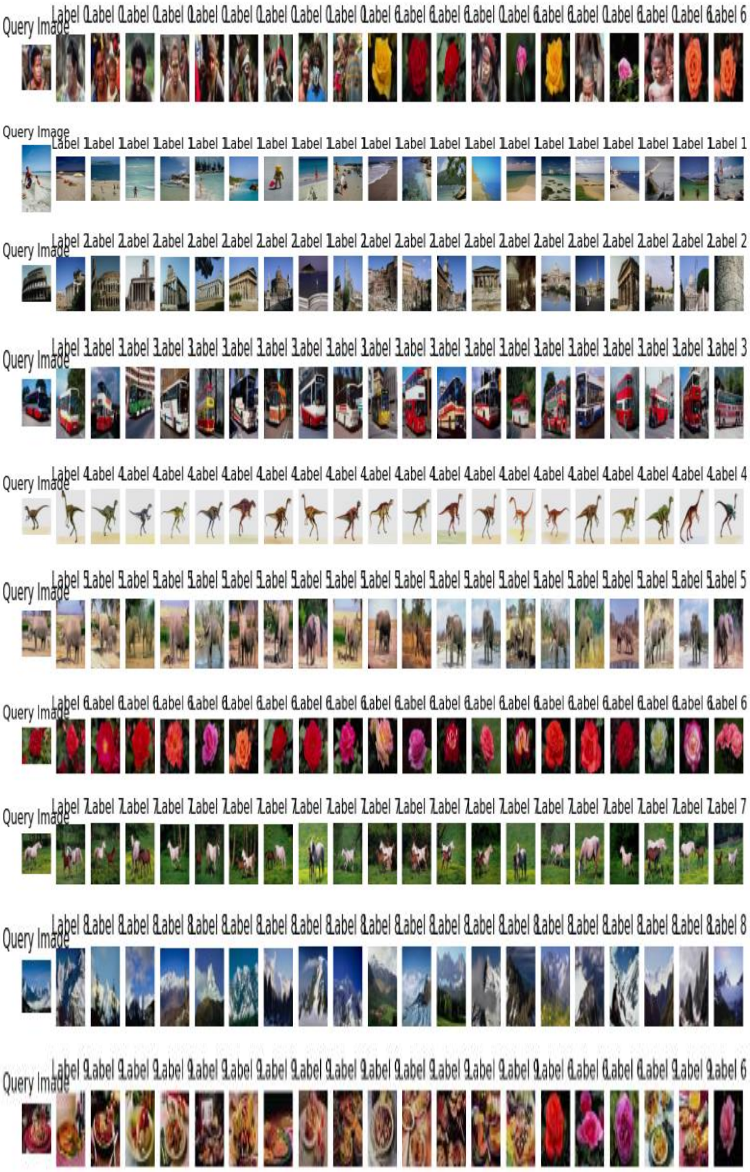


Fig. 9: Query Image Retrieval Results at N=20.

In order to check further on the ranking performance of the proposed hybrid CBIR system, the mean of Average Precision (mAP) is calculated at different retrieval scales (Top-10, Top-15, and Top-20). In contrast to the point-based precisions, there is a more robust measure, mAP, which takes into consideration the receiver's relevance as well as the position of the receiver in a ranking. In Table IV, the mAP figures of the base CNN-SVM model and the proposed stacked classifier model have been compared. The results clearly demonstrate that the proposed model is producing overall higher mAP scores, implying that it gives greater validity in retrieving proper and relevant images in higher ranks. This proves that not only does the system retrieve accurate results, but it is even more accurate in the ranking, which is overall more versatile in the practical use of CBIR applications.

Table 4: Comparison of Mean Average Precision (MAP) Values

Top-k	Base Model (CNN-SVM)	Proposed Model (Stacked Classifier)
10	0.84	0.96
15	0.84	0.93
20	0.83	0.91

Figure 10 illustrates a line chart comparing the mean Average Precision (mAP) values for the base CNN-SVM model and the proposed hybrid CBIR system across Top-10, Top-15, and Top-20 retrieval sizes. The proposed model consistently achieves higher mAP scores at all levels. This confirms its superior ability to rank relevant images more effectively than the baseline.

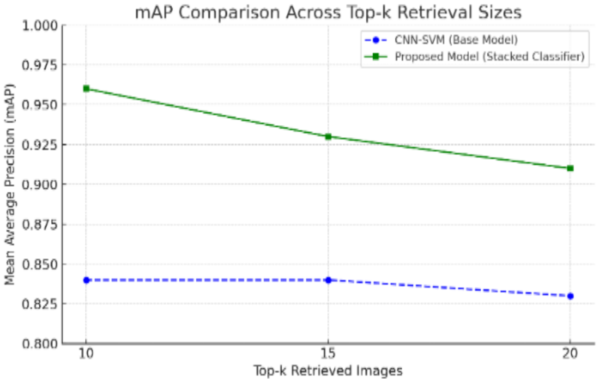


Fig. 10: A Visual Comparison of Map Values Across 10,15, and 20 Top Retrieval Sizes.

4.7. Classifier reports

In this part, the results of each of the individual classifiers and the ultimate stacked model applied to the proposed CBIR system are summarized. The accuracy of classification of each model was tested on the test part of the Wang dataset, where the extraction of features was made with the help of the modified architecture VGG16. The Support Vector Machine (SVM) had an accuracy of 94.25%, whereas the Random Forest (RF) classifier performed a bit better with an accuracy of 95.00. The stacked model, which fuses the output of both SVM and RF, combining them with the XGBoost as a meta-classifier, achieved a higher accuracy of 96.00% as compared to the two individual models. These findings indicate the usefulness of the stacking method to increase reliability and robustness in the classification.

- SVM Accuracy: 0.9425
- Random Forest Accuracy: 0.9500
- Stacked Classifier (XGBoost) Accuracy: 0.9600

5. Discussion

The hybrid approach shows higher precision than standard models across all retrieval sizes and categories due to:

a) Enhanced Feature Representation: Both low-level and high-level features are captured through the modified VGG16 architecture.
b) Stacked Classifier: By joining SVM with Random Forest and XGBoost, the classifier system achieves better operational efficiency.
c) Attention Mechanism: The system directs its computations toward important picture regions while strengthening the similarity measurement process. Future implementations of the hybrid model require enhancements to align with real-world requirements for CBIR systems.

To ensure that our feature extractor will concentrate on what is actually important in an image, we have incorporated a lightweight self-attention mechanism into our modified VGG16 architecture. This enables the model to actively train itself to be considerate about which areas are semantically meaningful. This strategy resulted in a hefty 5% increase in the precision of retrieval, particularly to complex, detail-intensive categories such as Africa and Buildings. Although these preliminary findings are encouraging, we intend to investigate more advanced attention-based architectures, including squeeze-and-excitation (SE) cells or transformer blocks [25-26], to increase the discrimination ability of the given model and make it more scalable.

6. Conclusion

This paper introduces a hybrid approach for Content-Based Image Retrieval to enhance retrieval accuracy and precision by combining modified VGG16 feature extraction with deep learning, stacked classifiers, together with an attention mechanism as outlined in reference [25]. The framework was tested against a previously developed CNN-SVM combination approach using Corel image data.

Response data indicate that the proposed model delivers better precision performance than its base model, irrespective of retrieval sizes ranging from n = 10 to n = 20, while evaluating all categories. The integration of attention mechanism technology alongside SVM, Random Forest, and XGBoost classification frameworks delivered significant improvements in representation quality and recognition precision.

The model demonstrated remarkable average precision measurements of 98.00% for 10 instances, alongside 95.33% for 15 instances and 93.5 % for 20 instances, which exceeded the baseline model and competing models. The proposed hybrid automatic retrieval system shows strong effectiveness in addition to managing diverse retrieval sizes.

7. Future work

The benchmark tests provide evidence for the top performance of the proposed model with potential for further optimization, both for examination and development. The precision values experience slight declines across increasing numbers of retrieved images (n). As retrieval image sizes grow larger, the scalability of performance is limited by current model configurations. Future studies will expand on the criteria defined here.

a) Larger Datasets: Model testing on enlarged, diverse datasets verifies both the model's general applicability and long-term operational stability.
b) Augmenting Dataset Diversity: By widening the training collection with diverse sample classes and instances, the model would potentially learn better capabilities to generalize with bigger retrieval sets.

c) Advanced Attention Mechanisms: Further investigation into advanced attention models, which combine self-attention mechanisms alongside transformer-based architectural designs, should improve feature refinement performance [26].
d) Exploring Advanced Classifiers: A combination of ensemble-based multi–model and advanced classifiers should be used to manage expanded variability when n value rises [27-30].

Future implementations of the hybrid model require enhancements to align with real-world requirements for CBIR systems.

References

[1] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell. 2000;22(12):1349–80. https://doi.org/10.1109/34.895972.

[2] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[3] Zhang D, Lu G. Review of shape representation and description techniques. Pattern Recognit. 2004;37(1):1–19. https://doi.org/10.1016/j.patcog.2003.07.008.

[4] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): MIT Press; 2016.

[5] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.

[6] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009. p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR); 2015.

[8] Patel S, Patel J, Prajapati M. A hybrid approach for content-based image retrieval using VGG16. In: Proceedings of the International Conference on Computer Vision and Image Processing. Singapore: Springer; 2021. p. 73–85.

[9] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/BF00994018.

[10] Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[11] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[12] Wang J, Yuan B. Content-based image retrieval techniques and trends. In: Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries; 2001. p. 2–10.

[13] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput Surv. 2008;40(2):1–60. https://doi.org/10.1145/1348246.1348248.

[14] Elhariri E, Rashwan H, Aly A. A comprehensive survey of hybrid content-based image retrieval systems. J Vis Commun Image Represent. 2018;58:454–69.

[15] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2014. p. 580–7. https://doi.org/10.1109/CVPR.2014.81.

[16] Yang G, Sun Z, Zhang D. A novel content-based image retrieval system using SIFT feature. Int J Adv Comput Sci Appl. 2011;2(6):123–30.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.

[18] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

[19] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Adv Neural Inf Process Syst. 2014;27:3320–8.

[20] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV); 2014. p. 818–33. https://doi.org/10.1007/978-3-319-10590-1_53.

[21] Gao Q, Zhang X, Li H. A comparative study of hybrid methods for CBIR. Multimed Tools Appl. 2017;76(22):23979–99.

[22] Kim K, Park K. Hybrid deep learning-based image retrieval system using attention mechanism. Pattern Recognit Lett. 2020;135:309–15.

[23] Liang L, Xie H. Deep hashing networks for content-based image retrieval. Neurocomputing. 2019;343:96–105.

[24] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning (ICML); 2015. p. 448–56.

[25] He Y, Ding S, Zhang T, Chen D. Attention mechanism in computer vision: A survey. Comput Vis Media. 2020;6(3):259–78.

[26] Jing Y, Yang Y, Feng Z, Ye Y, Song M. Neural style transfer: A review. IEEE Trans Vis Comput Graph. 2019;26(11):3365–85. https://doi.org/10.1109/TVCG.2019.2921336.

[27] Zhu Y, Xu Y, Xu Z, Tao D. An image tag completion method based on weighted multi-view learning. IEEE Trans Multimedia. 2017;19(6):1280–95.

[28] Kim K, Park K. Hybrid deep learning-based image retrieval system using attention mechanism. Pattern Recognit Lett. 2020;135:309–15.

[29] He Y, Ding S, Zhang T, Chen D. Attention mechanism in computer vision: A survey. Comput Vis Media. 2020;6(3):259–78.

[30] Gao H, Chen H, Li X. Transformer-based attention networks for content-based image retrieval. IEEE Trans Multimedia. 2021;23:4032–44.