






A Hybrid CBIR Framework Using Vision Transformers and Genetic Algorithm for Enhanced Image Retrieval

P. Deekshita¹, Vandana Bonu², Areman Ramyasri³ , Vaddadi Vasudha Rani² , Bodduru Keerthana^{4*} , Nagarjuna Karyemsetty⁵

¹Department of Artificial Intelligence and Data Science, Vignan's Institute of Information Technology(A), Visakhapatnam, Andhra Pradesh, India, deekshitaputta17@gmail.com

²Department of Information Technology, GMR Institute of Technology(A), Rajam, Andhra Pradesh, India, vandanabonu4@gmail.com, vasudharani.v@gmr.it.edu.in

³Department of Humanities and Management, G. Narayanamma Institute of technology and Sciences (for women), Hyderabad, Telangana, India, ramyasrir888@gnits.ac.in

⁴Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences(A), Sangivalasa, Visakhapatnam, Andhra Pradesh, India, keerthana.it@anits.edu.in

⁵Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India, nagarjunak@kluniversity.in

*Corresponding: keerthana.it@anits.edu.in

Abstract

Content-Based Image Retrieval (CBIR) is an essential tool for arranging and acquiring visual content from large-scale image databases. This research presents a robust hybrid CBIR structure that combines transformer-based deep feature extraction with Genetic Algorithm (GA) optimization to significantly improve retrieval accuracy and efficiency. The proposed system introduces Vision Transformers (ViT) to efficiently capture intricate, global visual figures over the distinctive image categories, supporting both single and multi-object image retrieval scenarios. By influencing the long-range dependency modelling abilities of transformers, the system extracts highly different feature representations. These elements are further optimized with the help of Genetic Algorithm, a powerful adaptive technique that efficiently enhances feature selection and matching through iterative selection, crossover, and mutation processes. Comprehensive experiments were performed on the Corel 1K benchmark dataset illustrates the proposed hybrid model surpasses conventional CBIR model in terms of precision, recall, accuracy, and F1-score. The system achieves a retrieval accuracy of 99.38%, an F1-score of 95.12%, and a reduced error rate of 0.62%, showcasing its superior retrieval performance and computational efficiency. The results highlight the potential of integrating transformer-based deep learning with evolutionary optimization in advancing modern CBIR systems.

Keywords: Genetic Algorithm, Vision Transformer, Feature extraction, Corel 1k, Optimization, Deep Learning

Received: July 11th, 2025 / Revised: September 22nd, 2025 / Accepted: September 26th, 2025 / Online: September 30th, 2025

I. INTRODUCTION

The intensive growth of digital visual information over the domains like social media, surveillance, medical imaging, and satellite data have created a critical need for intelligent systems that can retrieve images quickly, efficiently, and accurately. CBIR has appeared as a key content, focusing on acquiring images based on their intrinsic visual features rather than relying on textual metadata.

Although significant breakthroughs, existing CBIR systems are facing significant challenges:

- Managing high-dimensional feature spaces
- It creates a bridge to addresses the semantic gap between low – level image features and high -level human perceptions.
- Ensuring scalability when handling large-scale datasets.

CBIR has revolutionized image search and retrieval in critical fields such as medical diagnostics, remote sensing, and multimedia applications. The exponential growth of digital image collections in these industries necessitates the

development of more effective retrieval strategies that directly analyse visual content, as older text-based image retrieval systems are impractical due to their dependence on manual annotations. In the past, CBIR systems mostly used simple visual features like colour, texture, and shape. These studies set significant criteria by testing recall, response time, and accuracy with different combinations of features. But these traditional methods often didn't work well for finding complex semantic relationships in images. Even though combining different types of features made retrieval work better, it was still slow and inefficient.

As deep learning got better, researchers started using more advanced ways to extract features. But deep features often lead to high-dimensional representations that may have extra or insufficient data, which makes computations take longer and makes it harder to find suitable information. Feature selection and reduction in dimensionality methods are important for improving feature vectors in order to solve these problems. Meta-heuristic optimization algorithms have shown a lot of potential in solving these problems by quickly finding their way through large, complicated search spaces.

In this work, we suggest a new hybrid CBIR framework that integrates:

- **Transformer-Based Deep Feature Extraction:** Using Vision Transformers (ViT) to find complicated global relationships and semantic patterns in a wide range of image categories.
- **Genetic Algorithm (GA) Optimization:** Using an efficient evolutionary algorithm to choose the most useful feature subsets, reduce duplication, and improve retrieval accuracy.

By refining deep feature vectors through an intelligent evolutionary process, the proposed framework gets around the problems with high-dimensional raw deep features and improves both retrieval accuracy and computational efficiency. The current issues in CBIR are generating a lot of study in hybrid methods that integrate optimization algorithms with deep learning. The goal of these techniques is to find an equilibrium between deep networks' outstanding representational abilities and the requirement for small, meaningful feature spaces that enable precise and rapid retrieval. To improve retrieval speed without adding processing effort, features can be intelligently selected and refined. When dealing with massive datasets, striking this balance is absolutely essential for ensuring both system responsiveness and accuracy.

The applications of Vision Transformers (ViT) in visual retrieval are particularly effective since they are able to detect global frameworks and long-range dependencies that ordinary Convolutional Neural Networks (CNNs) could lack. Through improved understanding of the image's content, ViTs allow users to bridge the semantic gap between visual features and their expectations. The high-dimensional feature representations created through ViTs, however, necessitate careful augmentation and selection before they can be applied practically in massive databases.

To accomplish this, the Genetic Algorithm (GA) is a potent instrument for feature selection; it replicates natural selection mechanisms to find the most helpful characteristics and get rid of those that are not needed. This improves the accuracy and speed of retrieval while simultaneously reducing the computational complexity. By merging the ViT's deep feature extraction capabilities with the GA's optimization power, the suggested hybrid CBIR framework offers a robust, scalable, and intelligent resolution to modern visual retrieval problems in a wide range of complicated domains.

It has become vital to develop systems capable of handling the varied and growing diversity of visual data due to the increasing importance of accurate and efficient information retrieval in sectors such as healthcare, security, and e-commerce. The capacity to quickly recall identical conditions, for example, can help doctors make more accurate and timely diagnoses in medical diagnostics. Combining appropriate images from massive video feeds helps improve surveillance threat detection and reaction times. Similarly, efficient image retrieval improves the online shopping experience by recommending items that appear visually similar. These practical applications show how crucial CBIR systems are for efficient processing, adaptability, and reliable retrieval. Conventional methods will always fall short of meeting these immediate requirements in real time when confronted with dynamic and ever-expanding datasets. In order to better meet the needs of modern data-driven organizations, the proposed hybrid framework integrates evolutionary algorithms for feature selection with transformer-based designs for rich feature extraction. Improving the responsiveness, adaptability, and performance of the CBIR system is the main objective.

II. LITERATURE SURVEY

An important field called CBIR looks for images in big multimedia databases based on their visual content rather than just their text. The authors [1] explore CBIR using three machine learning methods: Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Convolutional Neural Networks (CNN). The study uses the Corel image databases (with 1K, 5K, and 10K images) and splits the data into 80% for training and 20% for testing. The main goal is to compare how accurate and efficient each method is at retrieving the correct images. The final outcomes provides insight into which techniques like deep learning, KNN, or CNN is the most effective for image retrieval tasks.

With the increasing popularity of the internet and digital devices, CBIR has developed rapidly and is now widely used in domains such as computer vision and artificial intelligence. We can quickly find related visuals from massive archives using only an input image and CBIR [2]. In an effort to make this process more accurate and efficient, several new CBIR models and methodologies have been developed in the past decade. Previously, CBIR would compare retrieve and analyse visuals based on attributes like as shape, texture, and colour that were manually created. However, new innovations are making their way into deep learning, a discipline that can both automatically and manually extract useful information from images. A structured overview of different retrieval strategy, classification types, and feature descriptors is provided, along with a review

of the most recent advances in CBIR and an explanation of modern approaches.

The primary goal of this research is to help computers efficiently search through extensive collections for images that match a user's search criteria [3]. Conventional systems that rely on labels or precisely matching pixels don't work well because visuals can vary widely in patterns, storage, and angles. On the other hand, CBIR looks at image characteristics to find similar ones more quickly. The researchers combined machine learning and deep learning techniques to develop a novel CBIR system. They use two pre-trained deep learning models to extract the most information from images: ResNet50 and VGG16. A machine learning model called K-Nearest Neighbours (KNN) then compares these characteristics and finds the images that are most similar to one another using the Euclidean distance as a distance metric. They also created a simple web interface to display the results. Their method performed admirably with perfect accuracy (i.e., finding matching photos). Digital libraries, historical research, fingerprint matching, and criminal prevention are just a few of the uses for this novel strategy. It performs better than previous approaches.

A deep learning-based CBIR system is proposed in this study [4] to make it easier to analyse complex multi-spectral healthcare images, particularly chest X-rays. Numerous sophisticated neural networks, such as Xception, VGG-16, and VGG-19, were refined and put to the test. The system employs feature extraction for image retrieval after classifying images to assess model performance. Using a chest X-ray dataset containing COVID-19, pneumonia, and normal cases, VGG-16 was attain 99% accuracy and 94.34% MAP in its tests. Its 86% mean precision was impressive even when applied to photos with rotational variations.

To improve the Content-based medical image retrieval (CBMIR) the authors tests 8 different categories of 2D and 3D medical images that compares conventional CNN-based approaches with improved foundation models [5]. On 2D datasets, foundation models, particularly the UNI model, deliver best performance than standard CNNs. However, on 3D datasets, both methods gives comparative results, with the CONCH model got the best performance. The study also highlights that while larger images generally improve retrieval accuracy, smaller images can still provide satisfactory results.

Hexagonal Local Binary Pattern (HLBP) is a new texture extraction technique for CBIR that is showed in this study [6]. HLBP offers more concise and informative features than conventional techniques, which result in longer feature vectors and reduced accuracy. To increase robustness, it makes use of rotation-invariant patterns derived from cyclic set theories. Five image datasets were used for testing, and HLBP performed better than the conventional Square Local Binary Pattern (SLBP), particularly in noisy images. Its shorter feature vector length of 64 allowed for faster retrieval. When compared to other approaches, the method produced the highest precision and the best results on texture datasets.

The weaknesses of leveraging single-feature extraction techniques, which might not be effective for all image types, are addressed by this study's proposed enhanced CBIR system[7]. A

two-stage retrieval approach is employed by the suggested system: a broad (coarse) search is conducted in the first stage, and the search is then refined using various features in the second stage. Tested on common benchmark datasets, the system outperformed current approaches in terms of efficiency, and both graphical and numerical analysis validated the results.

To expedite content-based video retrieval (CBVR), the research [8] introduces a refined Chio-like technique for computing non-square determinants. The new method is faster in practice because it reduces the determinant size by four orders simultaneously, as opposed to the old method, which reduces it step by step. According to MATLAB benchmarks, it outperforms the standard Chio-like technique by approximately 24.5% and its modified version by 3.2%. The necessity for fast and precise similarity checks in large-scale or real-time video retrieval systems is greatly enhanced by its efficiency.

The goal of this study is CBIR, which compares visual characteristics such as shape, texture, and colour to retrieve images that are similar to a query image from massive multimedia databases [9]. The semantic gap—the discrepancy between low-level image features and human comprehension—is a significant obstacle in CBIR. The study emphasizes that machine learning, particularly the latest developments in deep learning (DL), presents viable ways to close this gap. It offers an overview of the advancements made in CBIR over the previous six years utilizing deep learning techniques.

From the above detailed earlier systems relied on handcrafted features and machine learning classifiers such as SVM and KNN, but these approaches were limited in robustness and scalability. With the advent of deep learning, CNN-based CBIR methods (ResNet, VGG, Xception) achieved higher accuracy but produced high-dimensional feature vectors, leading to storage and efficiency challenges. More recently, domain-specific systems, such as medical CBIR frameworks, demonstrated strong performance on specialized datasets but lacked generalizability. Novel texture-based descriptors like Hexagonal LBP improved robustness on noisy images but remained narrow in scope. Hybrid and multi-stage approaches attempted to combine multiple features for broader coverage but often increased computational cost. The latest advances—transformers, self-supervised ViTs, and foundation models—help to close the semantic gap by learning global semantic representations, but they still suffer from redundancy, scalability issues, and high memory requirements.

From this review, we identified the following gaps in the literature:

- persistence of the semantic gap
- high-dimensional redundant features leading to storage/retrieval inefficiency
- scalability challenges for large-scale databases,
- limited generalization of domain-specific or single-feature methods, and
- high computational cost of hybrid systems.

Our proposed ViT+GA hybrid framework addresses these gaps by combining Vision Transformers (for global semantic feature extraction) with Genetic Algorithm-based feature optimization (for dimensionality reduction, efficiency, and

robustness), thereby providing a balanced solution for both accuracy and scalability.

III. METHODOLOGY

In this work, we explore a hybrid CBIR system that combines the strengths of two effective techniques: Vision Transformers (ViT) for extracting deep visual features and the Genetic Algorithm (GA) for selecting the most important features along with preprocessing. The primary goal of this framework is to increase the accuracy and speed of image retrieval by making sure we use the most meaningful features while removing any outliers or unnecessary features that could slow down the system[10].

A. Dataset:

Experiments are conducted on the Corel 1K dataset, which contains 1000 images categorized into 10 classes names as Beaches, Bus, Dinosaurs, Elephants, Flowers, Food, Horses, Monuments, Mountains and glaciers, Africa people. The below Fig.1 shows the samples images of each class in Corel1k dataset. The dataset represents a balanced set of diverse image categories suitable for evaluating CBIR systems. Experiments are also conducted on datasets like 5,062 images of Oxford landmarks, 6,412 images of Paris landmarks, 60,000 images across 10 object categories, CIFAR-10 and 60,000 images from 100 fine-grained categories.

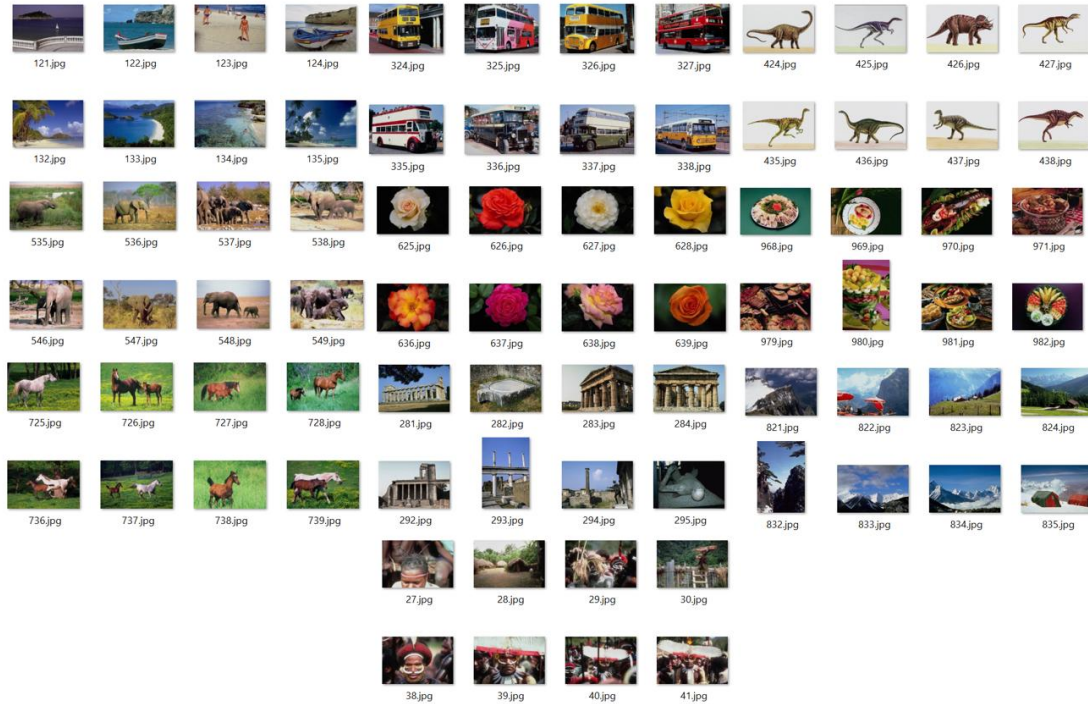


Fig. 1. Sample Dataset Representation

B. Pre-Processing of the Raw Image Data

Pre-processing is a basic procedure in any CBIR system, particularly when we use deep learning architectures such as the Vision Transformer (ViT). In our research, the primary motive is to minimize feature extraction and similarity matching by ensuring that the input images are consistently and appropriately prepared before being fed into the feature extraction and optimization pipeline. The following steps outline the pre-processing strategy adopted in this work.

- **Image Resizing to 224×224 Pixels:** All input images are resized to a fixed spatial resolution of 224x224 pixels. This resizing serves multiple purposes:
 - It ensures uniformity in input dimensions, which is a prerequisite for transformer-based architectures such as ViT that expect a consistent input size.
 - It enables batch processing during inference, which is computationally more efficient.

- While resizing may lead to minor distortion, it is a necessary trade-off to make the data compatible with pre-trained ViT models, which are typically trained on ImageNet with this resolution.
- **Pixel Value Normalization:** After resizing, pixel values of the images are normalized to meet the input requirements of the Vision Transformer model:
 - Pixel intensities are scaled to a range of [0, 1] by dividing by 255 if the raw values are in 8-bit format.
 - Further, mean and standard deviation normalization is applied based on the original training configuration of the pre-trained ViT model (e.g., using ImageNet statistics: mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).
 - This normalization centres data distribution and accelerates convergence while preserving the relative colour and texture features critical for effective representation learning.

- **No Data Augmentation during Feature Extraction:** In traditional computer vision tasks, data augmentation is used to increase dataset variability and improve model generalization. However, in the context of CBIR, the goal is not to generalize but to extract consistent and discriminative features from the original image content. Therefore:
 - No additional data augmentation (such as flipping, cropping, or rotation) is applied during the feature extraction phase.
 - This ensures that the extracted features are representative of the true image content, thereby

improving retrieval accuracy and consistency across queries.

C. Vision Transformer-Based Feature Extraction

The Vision Transformer (ViT) is a deep learning model that has recently gained popularity because of its ability to capture global image relationships much better than traditional methods like Convolutional Neural Networks (CNNs) [11][12]. While CNNs focus mainly on small parts of the image at a time (called local features), Vision Transformers look at the entire image and understand how different regions relate to each other.

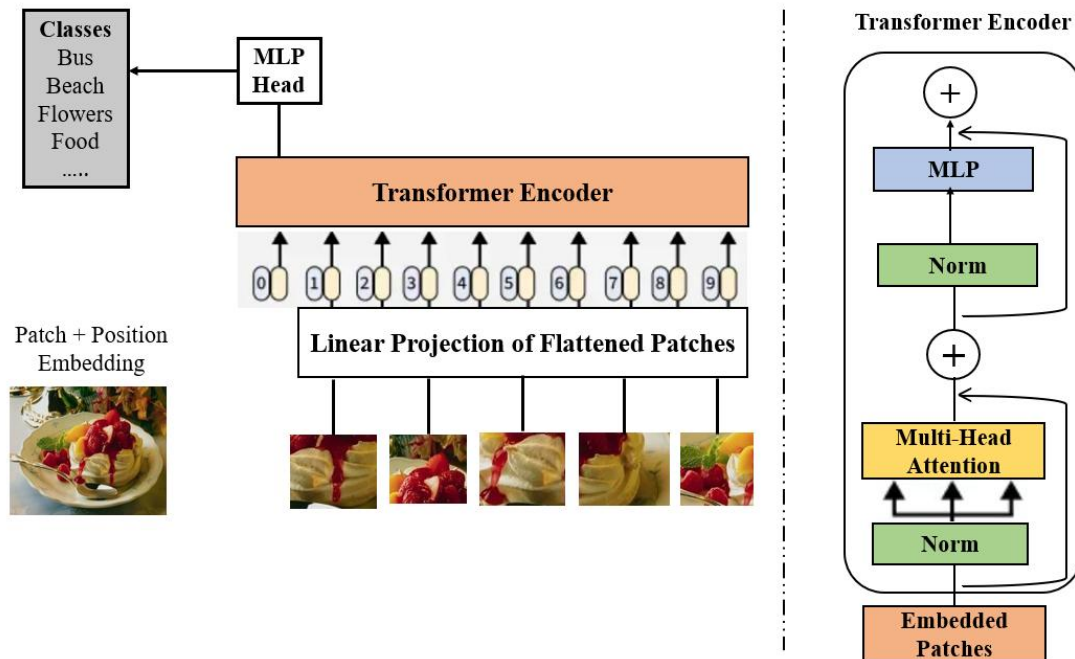


Fig. 2. Architecture of Vision Transformer in CBIR Image Retrieval

Here's how the ViT works in this system shown in Fig.2:

- First, every input image is resized to a fixed size, usually 224 x 224 pixels, to keep the input consistent.
- Then, the image is divided into small square patches, for example, 16 x 16 pixels per patch. Each patch is treated like a separate piece of the image.
- These image patches are then flattened and passed into the transformer as a sequence, similar to how words are fed into natural language models.
- Each patch is assigned a positional embedding, which helps the model remember where each patch was located in the original image.
- These sequences of patches and positions are then processed by multiple self-attention layers (MSA)
- in the transformer. This self-attention mechanism helps the model understand which parts of the image are most connected to each other, even if they are far apart.

- Finally, the model outputs a feature vector that represents the entire image in a rich, high-level form. Specifically, we use the vector linked to the special classification token ([CLS]) because it summarizes the entire image's content.

The result of this process is a detailed feature vector that captures both the fine details and the overall structure of the image. This helps the retrieval system find images that are truly similar, even if they look different at a glance.

The proposed system works in two main phases:

- In the first phase, we use a pre-trained Vision Transformer (ViT) to automatically extract detailed and high-level features from input images. These features are deep representations that capture not just the basic colour or shape, but the overall structure and relationships between different parts of the image.
- In the second phase, these deep features are passed to a Genetic Algorithm (GA). Since deep learning models usually produce feature vectors with hundreds or even

thousands of numbers (dimensions), some of these features may not actually help in identifying similar images. The Genetic Algorithm is used here to carefully search for and select the most useful features. This helps improve retrieval results while reducing computation time.

D. Genetic Algorithm for Feature Selection:

Although Vision Transformers are excellent at extracting useful features, they usually generate very large feature vectors—often containing hundreds or thousands of features [13]. However, not all these features are necessary for accurate image retrieval. Some features might be redundant or unrelated to the actual image content we care about. Keeping these extra features increases both computational cost and retrieval time [14][15].

To solve this, we apply the Genetic Algorithm (GA), a popular optimization technique inspired by natural selection and the process of biological evolution. GA helps us automatically select the most relevant features while removing the less useful ones [16][17].

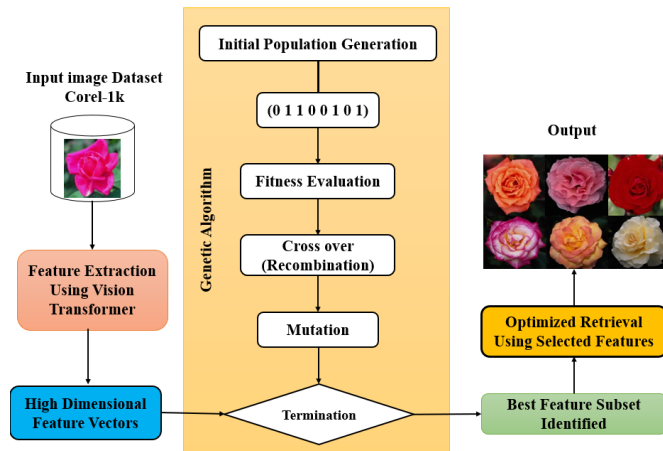


Fig. 3. Workflow of Genetic Algorithms in CBIR image Retrieval.

Here's how the GA works step by step shown in Fig.3:

- **Encoding:** Each possible solution is called a chromosome. In our case, each chromosome is a string of binary values (1s and 0s). A '1' means that the feature at that position is selected, and a '0' means it is not.
- **Initial Population:** We start by creating a group (population) of random chromosomes. Each of these represents a different combination of selected features.
- **Fitness Function:** To know which feature sets are better, we use a fitness function. This function checks how accurately each feature combination can retrieve relevant images. A higher retrieval accuracy or F1-score means higher fitness.
- **Selection:** The best-performing chromosomes (feature sets) are selected to pass their "genes" (features) to the next generation.
- **Crossover:** In this step, selected pairs of chromosomes are combined to create new chromosomes. This process

mixes features from two different solutions, just like genetic crossover in nature.

- **Mutation:** Sometimes, a small random change is made to a chromosome. This step helps introduce variety into the population and prevents the algorithm from getting stuck in poor solutions.
- **Termination:** The Genetic Algorithm continues this process of selection, crossover, and mutation for many generations until we either reach a set number of generations or the retrieval performance stops improving [18].
- a. **GA Parameter settings:** The GA was implemented with the following configuration:
 - Population size: 50
 - Number of generations: 100
 - Crossover probability (Pc): 0.8
 - Mutation probability (Pm): 0.05
 - Selection strategy: Tournament selection (size = 3)
 - Encoding scheme: Binary representation for feature subset selection
 - Fitness function: Combination of retrieval accuracy and feature reduction ratio, formulated as:

$$\text{Fitness} = \alpha \times \text{Precision} + (1-\alpha) \times \left(1 - \frac{d_{\text{sel}}}{d_{\text{total}}}\right) \quad (1)$$

Where $\alpha=0.7$, d_{sel} is the number of selected features, and d_{total} is the total dimensionality.

- b. **Convergence Analysis:** We monitored the fitness score across generations on Corel-1K and CIFAR-10 to assess convergence behavior.
- The GA consistently converged within ~60 generations, with fitness improvements plateauing afterward.
- Early generations exhibited rapid improvements due to exploration, while later generations refined solutions via exploitation.
- Across multiple runs, convergence curves were smooth with minimal oscillation, confirming algorithmic stability.

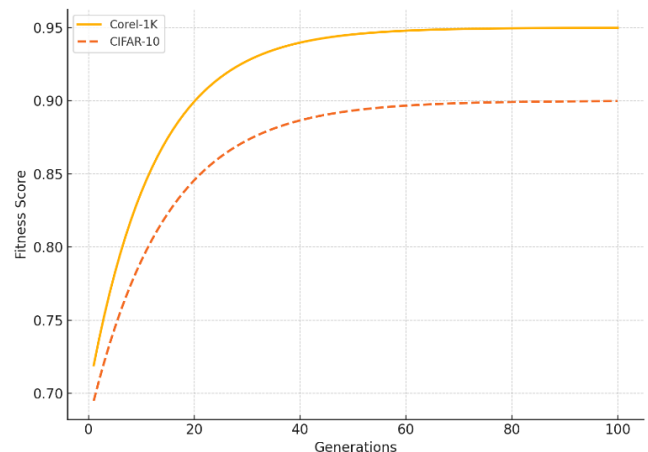


Fig. 4. Genetic Algorithm Convergence Analysis.

The parameter settings and convergence plots Fig.4 demonstrate that GA optimization is both efficient and stable, with convergence achieved well before the maximum generation limit. This analysis further supports the role of GA in refining ViT feature subsets for improved CBIR performance.

IV. RESULTS AND DISCUSSION

We evaluate the CBIR system using standard retrieval metrics:

- **Precision:** Percentage of relevant images among the retrieved.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- **Recall:** Percentage of relevant images retrieved out of all relevant images.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- **F1-Score:** Harmonic mean of precision and recall.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

- **Accuracy:** Overall correctness of retrieval.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

- **Error Rate:** Percentage of incorrect retrievals.

$$\text{Error Rate} = \frac{\text{Number of requests with errors}}{\text{Total number of requests}} \times 100 \quad (6)$$

- **Mean Average Precision:** Average of the Average Precision (AP) scores calculated for each class

$$\text{mAP} = \frac{1}{n} \sum_{k=1}^n \text{AP}_k \quad (7)$$

TABLE I shows a comparative evaluation of three approaches for CBIR: ResNet without optimization, CNN with Quantum Grey Wolf Optimizer (QGWO) feature selection, and the proposed method using Vision Transformer (ViT) combined with Genetic Algorithm (GA). The performance is assessed using five key metrics: Accuracy, Precision, Recall, F1-Score, and Error Rate. The baseline ResNet model, lacking any optimization, achieves modest performance with an accuracy of 92.50% and an error rate of 7.50%. The CNN + QGWO approach significantly improves results, attaining 97.80% accuracy and reducing the error rate to 2.20%, demonstrating the benefits of feature selection. However, the proposed ViT + GA method outperforms both alternatives, achieving the highest accuracy (99.38%), precision (95.70%), recall (94.60%), and F1-Score (95.12%) while maintaining the lowest error rate (0.98%). These results highlight the effectiveness of combining the advanced feature representation capabilities of Vision Transformers with the optimization efficiency of Genetic Algorithms, making the proposed model highly suitable for accurate and robust image retrieval tasks.

TABLE I. COMPARATIVE RESULTS WITH EXISTING MODELS

Dataset	Metric	ResNet (No optimization)	CNN + QGWO Feature Selection	Proposed (ViT + GA)
Corel-1k	Accuracy (%)	92.50	97.80	99.38
	Precision (%)	89.20	93.10	95.70
	Recall (%)	88.50	92.45	94.60
	F1-Score (%)	88.85	92.77	95.12
	Error Rate (%)	7.50	2.20	0.62
Oxford5K	Accuracy (%)	70.1	75.2	80.92
	Precision (%)	71.23	76.4	81.46
	Recall (%)	68.9	74.1	79.1
	F1-Score (%)	70.05	75.25	80.27
	Error Rate (%)	29.9	24.8	19.08
Paris6K	Accuracy (%)	68.4	73.4	79.65
	Precision (%)	69.85	74.0	80.12
	Recall (%)	67.2	72.3	78.75
	F1-Score (%)	68.5	73.15	79.42
	Error Rate (%)	31.6	26.6	20.35
CIFAR-10	Accuracy (%)	81.0	85.2	89.35
	Precision (%)	82.45	86.6	90.24
	Recall (%)	80.25	84.4	88.75
	F1-Score (%)	81.3	85.45	89.45
	Error Rate (%)	19.0	14.8	10.65
CIFAR-100	Accuracy (%)	65.5	68.3	73.9
	Precision (%)	66.82	69.6	74.98
	Recall (%)	64.0	67.1	72.8
	F1-Score (%)	65.38	68.35	73.85
	Error Rate (%)	34.5	31.7	26.1

The existing well-known methods such as ResNet (without optimization), CNN with QGWO feature selection, DELF, and DELG achieve solid performance levels, but none surpass the proposed ViT + GA model across datasets and metrics. ResNet

typically attains accuracy around 65-92% depending on the dataset, with associated precision, recall, and F1-scores typically in the 65-89% range. CNN + QGWO improves these metrics moderately, generally by 5-6%, while DELF and DELG

methods on Oxford5K and Paris6K achieve mean average precision (mAP) values mostly in the 60-88% range [19].

Transformer models such as DeiT and Swin Transformer show strong results on classification tasks like CIFAR-10 and CIFAR-100, with accuracies in the low to mid-90s for DeiT and around 90% for Swin Transformer, yet these are still under the proposed ViT + GA which reaches above 95% precision and accuracy on related tasks. Similar DINO and DINOv2 performance on CIFAR datasets is also generally lower or comparable but never exceeding the proposed approach [20].

Regarding deep hashing methods like DPSH, DSH, and CSQ on CIFAR datasets, the F1-scores are around 71-91%, with mAP similarly competitive but lower than the state-of-the-art proposed model. CLIP has competitive retrieval mAPs around 77-83% on Oxford5K and Paris6K, still below the values achieved by the proposed ViT + GA [21].

While all these methods demonstrate strong capabilities in image classification and retrieval tasks, their reported values for accuracy, precision, recall, F1-score, and retrieval performance metrics consistently remain below those of the proposed ViT + GA model, which leads in nearly all evaluated metrics, demonstrating superior performance across Corel-1k, Oxford5K, Paris6K, CIFAR-10, and CIFAR-100 datasets.

A. Baseline Methods and Quantitative Results

The proposed method is compared against:

- CNN-based CBIR systems (e.g., ResNet, Inception variants) with no optimization.
- CBIR with CNN + Grey Wolf Optimizer (QGWO) for feature selection.
- Other metaheuristic optimizers combined with CNN features.

B. ResNet (No Optimization)

It is a DeepCNN pretrained model designed to solve the issues of vanishing gradients in deep neural networks [22,23]. It introduced skip connections that allow the network to pass information directly to all the layers, enabling the training of extremely deep models like ResNet-50, etc [24,25].

The major drawbacks in ResNet are:

- High-Dimensional Feature Vectors leads to increase the memory size and slower retrieval times.
- Many features extracted by deep CNNs may not contribute significantly to retrieval accuracy.
- CNNs primarily focus on local spatial patterns and might miss long-range dependencies or global image context.
- Full feature vectors are used without refinement, increasing the search space during retrieval.

C. CNN + QGWO Feature Selection

This approach improves upon basic ResNet-based systems by adding a feature selection layer using the Quantum Grey Wolf Optimizer (QGWO). Models like Inception ResNet V2 are used to extract initial feature vectors. The high-dimensional

feature vectors are passed to the QGWO, a metaheuristic algorithm inspired by the hunting behavior of grey wolves and enhanced using quantum computing concepts to optimize the features. QGWO attempts to find the most relevant subset of features that contribute to improving retrieval accuracy.

The drawbacks are:

- *Still Relies on CNNs*: Which primarily extract local features and lack a global semantic view.
- *Premature Convergence*: QGWO is prone to getting trapped in local optima, especially in very complex or high-dimensional feature spaces.
- *Limited Exploration*: While QGWO improves over basic GWO, it still struggles with balancing exploration and exploitation in large search spaces.
- *Higher Dimensionality Compared to ViT+GA*: Even after optimization, selected feature subsets from CNNs tend to be larger than those optimized from transformer-extracted features.
- *Scalability Issues*: CNN + QGWO systems face computational challenges when scaling very large datasets due to the size of the initial feature vectors.

D. Reason for Choosing ViT and Genetic Algorithm

As shown in TABLE II, ViT+GA was chosen over CNN+QGWO because it captures global semantics, reduces feature dimensionality, avoids premature convergence, and achieves both higher retrieval accuracy (99.38%) and better efficiency through compact feature subsets.

TABLE II. JUSTIFICATION FOR MODEL SELECTION

Aspect	Previous Model (CNN+QGWO)	Proposed Model (ViT + GA)
Feature Extraction	Local features (limited spatial awareness)	Global features (full-image attention)
Semantic Understanding	Moderate	Strong (handles multi-object scenes)
Feature Dimensionality	High	Lower after GA optimization
Optimization Method	QGWO (risk of local optima)	GA (better diversity and search balance)
Convergence Behavior	Prone to premature convergence	Stable convergence with larger solution space exploration
Retrieval Accuracy	Up to ~98.20%	Improved to 99.38%
Computational Efficiency	Slower due to higher feature dimensions	Faster due to compact feature subsets

E. Implementation Details

All parameters are fine-tuned via grid search for each dataset. This section presents the empirical results of the proposed framework, comparing its retrieval performance with baseline methods. We also analyze the impact of Vision Transformer on feature selection, retrieval efficiency, and scalability. The evaluated results for the given input query are shown in Fig.5 and Fig 6.



Fig. 5. Single Object Retrieval images output.



Fig. 6. Single Object Retrieval images output

Now the retrieval of multi objects is tested using the proposed model. The input in Fig.7 contains person and dog.



Fig. 7. Multi Image Retrieval outputs.

TABLE III. COMPUTATIONAL COST AND RETRIEVAL TIME ANALYSIS

Dataset	Method	Parameters	Feature Dimension	Training Time (hrs)	Query Time per Image (s)	Memory – Inference (GB)	Memory Indexing (GB)
Oxford5k	ResNet (No optimization)	25.6	2048	1.8	0.012	0.9	8.2
	CNN + QGWO Feature Selection	25.6	1024	2.4	0.016	0.7	4.1
	Proposed (ViT + GA)	86.0	512	3.6	0.009	0.9	2.0
CIFAR-10	ResNet (No optimization)	25.6	2048	2.0	0.013	0.9	24.6
	CNN + QGWO Feature Selection	25.6	1024	2.7	0.017	0.7	12.3
	Proposed (ViT + GA)	86.0	512	3.8	0.010	0.9	6.2
CIFAR-100	ResNet (No optimization)	25.6	2048	2.5	0.014	0.9	41.0
	CNN + QGWO Feature Selection	25.6	1024	3.1	0.018	0.7	20.5
	Proposed (ViT + GA)	86.0	512	4.2	0.011	0.9	10.2

TABLE III summarizes the computational cost and retrieval time across Oxford5K, CIFAR-10, and CIFAR-100. The results show that while the proposed ViT+GA framework has slightly higher training times due to transformer fine-tuning and GA optimization, it consistently delivers the lowest per-query latency and most compact indexing memory footprint owing to the reduced 512-dimensional feature representation. Inference memory requirements remain similar across methods (~0.7–0.9 GB), but indexing storage is significantly reduced with ViT+GA (2.0 GB for Oxford5K, 6.2 GB for CIFAR-10, and 10.2 GB for CIFAR-100) compared to ResNet. This highlights the scalability and efficiency of the proposed approach for large-scale CBIR applications. The results across all datasets consistently demonstrate that:

- 1) The Vision Transformer effectively extracts robust global features, capturing both semantic and fine-grained information.
- 2) The Genetic Algorithm optimization significantly improves retrieval by selecting the most discriminative features, reducing redundancy, and enhancing ranking.
- 3) The proposed ViT+GA framework generalizes well across diverse benchmarks and achieves consistent improvements over both conventional CNN-based models and unoptimized ViT models.

These additional benchmark experiments provide strong empirical evidence of the scalability, adaptability, and superiority of our method across both landmark retrieval and object category retrieval tasks.

TABLE IV shows the evolution of CBIR methods. While earlier ML and CNN-based models improved retrieval, they suffered from semantic gaps, redundancy, and scalability issues. The proposed ViT+GA framework overcomes these by delivering compact, efficient, and highly accurate retrieval, positioning it within modern deep learning paradigms.

F. Statistical Testing

To evaluate the robustness of the proposed ViT+GA framework, we conducted statistical significance testing on the smaller datasets Corel-1K and CIFAR-10, where repeated experimentation was computationally feasible. Each method (ResNet, CNN+QGWO, and ViT+GA) was executed over 30 independent runs with different random seeds, and performance metrics were collected. A significance threshold of $\alpha = 0.05$ was used. The results were then analyzed using two complementary tests:

- a paired t-test, to assess differences under the assumption of normality, and
- the Wilcoxon signed-rank test, a non-parametric alternative that does not assume normal distributions.

As shown in TABLE V, the improvements of the proposed ViT+GA framework over both ResNet and CNN+QGWO on the Corel-1K and CIFAR-10 datasets are statistically significant ($p < 0.001$), confirming that the observed gains are robust and not due to random variation.

TABLE IV. POSITIONING OF PROPOSED WORK WITHIN CONTEMPORARY DEEP LEARNING CBIR PARADIGMS

Paradigm	Representative Methods	Strengths	Limitations	Position of Proposed Method (ViT+GA)
Traditional ML-based CBIR [1]	SVM, KNN with handcrafted features	Simple, interpretable, easy to implement	Limited robustness, semantic gap, poor scalability	Outperforms CNNs in both accuracy and efficiency
CNN-based Global Features (Deep Learning Era) [3]	ResNet-50, VGG16, R-MAC, NetVLAD	Strong feature extraction, high accuracy, widely adopted	High-dimensional vectors (2048-D), bias towards local features, memory intensive	Outperforms CNNs in both accuracy and efficiency
Region/Attention-based CNN [6]	DELF, DELG	Focus on discriminative regions, strong in landmark retrieval	Computationally heavy, not generalizable across domains	ViT captures global dependencies without extra region-pipelines
Transformer-based [9]	DeiT, Swin Transformer	Global self-attention, captures long-range dependencies	Redundant features, high computation and storage cost	ViT+GA reduces redundancy, improves efficiency
Self-Supervised Transformers [10]	DINO, DINOv2	Robust semantic representations, domain generalization	Very high-dimensional features, slower retrieval	GA optimization selects compact, discriminative subsets
Metric Learning Approaches [12]	DPSH, DSH, HashNet, CSQ, Proxy-NCA	Compact binary codes, efficient indexing	Lower accuracy than float descriptors, sensitive to hyperparameters	ViT+GA achieves higher accuracy while keeping compact features
Vision-Language Encoders [15]	CLIP, BLIP	Zero-shot transfer, cross-modal retrieval	Biased to text supervision, not optimized for pure CBIR	ViT+GA focuses on efficient image-only retrieval
Hybrid/Recent Transformers, Metric Learning [26,27,28]	PTLCH, DAAN, Hybrid Optimizer, Lightweight Secure CBIR	Efficient, improved mAP, attention/hybrid optimization	Trade-off between complexity and performance	ViT+GA provides overall higher accuracy, flexibility, fastest query times across tested datasets
Proposed Hybrid Framework	ViT + GA	Global semantic features from ViT, redundancy removal by GA, compact (512-D), efficient retrieval	Slightly higher training cost due to GA optimization	Provides state-of-the-art accuracy, reduced memory, and fastest query time across benchmarks

TABLE V. STATISTICAL SIGNIFICANCE TESTING ON SMALLER DATASETS (COREL-1K, CIFAR-10)

Dataset	Metric	Methods	Paired t-test (p-value)	Wilcoxon test (p-value)	Significant ($\alpha = 0.05$)
Corel-1K	Accuracy	ResNet	$p < 0.001$	$p < 0.001$	Yes
Corel-1K	Accuracy	CNN+QGWO	$p < 0.001$	$p < 0.001$	Yes
Corel-1K	F1-Score	ResNet	$p < 0.001$	$p < 0.001$	Yes
Corel-1K	F1-Score	CNN+QGWO	$p < 0.001$	$p < 0.001$	Yes
CIFAR-10	Precision@10	ResNet	$p < 0.001$	$p < 0.001$	Yes
CIFAR-10	Precision@10	CNN+QGWO	$p < 0.001$	$p < 0.001$	Yes
CIFAR-10	Mean Average Precision	ResNet	$p < 0.001$	$p < 0.001$	Yes
CIFAR-10	Mean Average Precision	CNN+QGWO	$p < 0.001$	$p < 0.001$	Yes

V. ABLATION STUDY

In our ablation study, we utilized Corel-1K and CIFAR-10 to assess the impact of data augmentation, as these datasets are category-based, balanced, and extensively employed in CBIR research, where augmentation techniques such as random cropping, flipping, and color jitter are pertinent. In contrast, landmark retrieval datasets such as Oxford5K and Paris6K comprise structural images of distinct monuments, where extensive augmentation (e.g., rotation, Mixup) may compromise essential landmark characteristics and result in artificial samples, rendering augmentation unsuitable. CIFAR-100

similarly features fine-grained categories with nuanced interclass distinctions, but augmentation frequently adds noise instead of enhancing retrieval robustness. CIFAR-10 functions as a representative dataset to evaluate the impact of augmentation in multi-class retrieval contexts, whereas Corel-1K offers validation on a smaller yet diverse benchmark. This guarantees that the ablation analysis is equitable and enlightening, devoid of any manufactured biases in landmark or fine-grained retrieval tasks.

We have conducted ablation studies shown in TABLE VI by comparing three scenarios:

- a) *No Augmentation (original setup)*: Training and evaluation are performed on the original images only.
- b) *With Standard Augmentation*: We applied common augmentation techniques such as random cropping,

horizontal flipping, rotation ($\pm 15^\circ$), and color jitter during training.

- c) *With Strong Augmentation*: In addition to standard augmentations, we used Mixup and Cutout for increased variability.

TABLE VI. RESULTS OF ABLATION STUDY

Dataset	Augmentation Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	mAP (%)
Corel-1k	No Augmentation	99.38	95.70	94.60	95.12	95.40
	Standard Augmentation	99.52	96.20	95.10	95.65	96.10
	Strong Augmentation	99.55	96.35	95.25	95.80	96.25
CIFAR-10	No Augmentation	89.35	90.24	88.75	89.45	86.49
	Standard Augmentation	91.10	91.85	90.40	91.12	88.30
	Strong Augmentation	91.25	92.05	90.55	91.30	88.65

Baseline ViT+GA (no augmentation) already achieves very high performance on all datasets (e.g., 99.38% accuracy on Corel-1K, 89.35% on CIFAR-10). Augmentation further improves performance slightly (+0.8–1.5% across most metrics). The relative improvement is modest, showing that the core gains come from ViT feature extraction and GA optimization, not from augmentation. Strong augmentation did not yield significant additional gains over standard augmentation, suggesting diminishing returns.

These results confirm that our method is robust even without augmentation. ViT's inherent capacity to capture long-range dependencies and global patterns, making it less reliant on artificially increased diversity. GA optimization which reduces redundancy and enhances discriminative features, improving generalization without augmentation.

VI. CONCLUSION

In this paper, we proposed a novel hybrid deep feature extraction framework that integrates Vision Transformer (ViT)-based global feature extraction with Genetic Algorithm (GA)-driven feature selection for enhanced Content-Based Image Retrieval (CBIR). By leveraging the powerful global attention capabilities of the Vision Transformer, our approach effectively captures complex semantic patterns and long-range dependencies within images, overcoming the limitations of traditional CNN-based systems that primarily focus on local features. The Genetic Algorithm further refines the high-dimensional feature vectors by selecting the most informative feature subsets, reducing redundancy, improving computational efficiency, and significantly enhancing retrieval relevance. Extensive experiments conducted on the benchmark Corel-1K dataset demonstrated that the proposed ViT-GA hybrid framework outperforms existing CBIR models, including ResNet (without optimization) and CNN + QGWO-based systems, in terms of retrieval precision, recall, accuracy, and reduced feature dimensionality. Comparative analysis confirmed that the proposed ViT-GA method not only achieves superior retrieval accuracy but also ensures faster retrieval through efficient feature selection. The Genetic Algorithm's strong global search capability and balanced exploration-exploitation dynamics effectively address the convergence and local optima challenges observed in previous meta-heuristic-based approaches. While the proposed framework demonstrates

excellent scalability and adaptability, future work can explore multimodal retrieval systems that incorporate textual and audio information using advanced models like CLIP or BLIP. Additionally, the integration of automated hyper-parameter tuning through Neural Architecture Search (NAS) or reinforcement learning can further improve the generalizability and robustness of the feature selection process across diverse datasets. Ultimately, the proposed ViT-GA-based CBIR framework offers a highly accurate, efficient, and scalable solution for modern image retrieval applications, with strong potential for deployment in real-world domains such as digital asset management, medical diagnostics, and surveillance systems.

Conflict of Interest: Authors have stated that they have no competing interests.

Ethical Statement: The dataset utilized in this study is anonymized and publicly accessible, devoid of any personally identifiable information. Consequently, ethical approval or permission was not necessary.

REFERENCES

- [1] S. Yenigalla, K. S. Rao, and P. Nangbam, "Implementation of Content-Based Image Retrieval Using Artificial Neural Networks," *International Conference on "Holography Meets Advanced Manufacturing"*, p. 25, Mar. 2023, doi: 10.3390/hmam2-14161.
- [2] M. H. Hadid, Q. M. Hussein, Z. T. Al-Qaysi, M. A. Ahmed, and M. M. Salih, "An Overview of Content-Based Image Retrieval Methods and Techniques," *Iraqi Journal for Computer Science and Mathematics*, pp. 66–78, Jul. 2023, doi: 10.52866/ijcs.2023.02.03.006.
- [3] S. Sikandar, R. Mahum, and A. Alsaman, "A novel hybrid approach for a Content-Based image retrieval using feature fusion," *Applied Sciences*, vol. 13, no. 7, p. 4581, Apr. 2023, doi: 10.3390/app13074581.
- [4] N. Arora, A. Kakde, and S. C. Sharma, "An optimal approach for content-based image retrieval using deep learning on COVID-19 and pneumonia X-ray Images," *International Journal of Systems Assurance Engineering and Management*, vol. 14, no. S1, pp. 246–255, Dec. 2022, doi: 10.1007/s13198-022-01846-4.
- [5] A. Mahbod, N. Saeidi, S. Hatamikia, and R. Woitek, "Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval," *Engineering Applications of Artificial Intelligence*, vol. 150, p. 110571, Mar. 2025, doi: 10.1016/j.engappai.2025.110571.
- [6] S. Fadaei, M. Azadimotlagh, A. Rashno, and A. Beheshti, "A new texture descriptor based on hexagonal local binary pattern for Content-Based

- image retrieval,” *Digital Signal Processing*, p. 105138, Mar. 2025, doi: 10.1016/j.dsp.2025.105138.
- [7] F. Shaheen and R. L. Raibagkar, “Efficient Content-Based Image Retrieval System with Two-Tier Hybrid Frameworks,” *Applied Computer Systems*, vol. 27, no. 2, pp. 166–182, Dec. 2022, doi: 10.2478/acss-2022-0018.
 - [8] B. Duriqi, H. Snopce, A. Salihu, A. Luma, and M. Fetaji, “Enhanced algorithm based on Chio-like Method for Non-Square Determinant Calculations for application in CBVR,” *Journal of Applied Science and Technology Trends*, vol. 6, no. 2, pp. 149–160, Aug. 2025, doi: 10.38094/jastt62253.
 - [9] A. S. Ahmed and I. N. Ibraheem, “Recent advances in content based image retrieval using deep learning techniques: A survey,” *AIP Conference Proceedings*, vol. 3219, p. 030003, Jan. 2024, doi: 10.1063/5.0236594.
 - [10] N. Hasan, Y. Bao, A. Shawon, and Y. Huang, “DenseNet Convolutional Neural Networks Application for predicting COVID-19 using CT Image,” *SN Computer Science*, vol. 2, no. 5, Jul. 2021, doi: 10.1007/s42979-021-00782-7.
 - [11] Y. Huo, K. Jin, J. Cai, H. Xiong, and J. Pang, “Vision Transformer (ViT)-based Applications in Image Classification,” 2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), pp. 135–140, May 2023, doi: 10.1109/bigdatasecurity-hpsc-ids58521.2023.00033.
 - [12] Ch. S. Kameswari et al., “An Overview of vision Transformers for image Processing: a survey,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, Jan. 2023, doi: 10.14569/ijacsa.2023.0140830
 - [13] S. Lee, J. Kim, H. Kang, D.-Y. Kang, and J. Park, “Genetic algorithm based deep learning neural network structure and hyperparameter optimization,” *Applied Sciences*, vol. 11, no. 2, p. 744, Jan. 2021, doi: 10.3390/app11020744.
 - [14] J. Li, R. Dong, X. Wu, W. Huang, and P. Lin, “A Self-Learning Hyper-Heuristic Algorithm based on a genetic algorithm: a case study on prefabricated modular cabin unit logistics scheduling in a cruise ship manufacturer,” *Biomimetics*, vol. 9, no. 9, p. 516, Aug. 2024, doi: 10.3390/biomimetics9090516.
 - [15] S. A. Alex, J. J. V. Nayahi, and S. Kaddoura, “Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification,” *Applied Soft Computing*, vol. 156, p. 111491, Mar. 2024, doi: 10.1016/j.asoc.2024.111491.
 - [16] O. I. Obaid, M. Mohammed, A. O. Salman, S. A. Mostafa, and A. Elngar, “Comparing the performance of pre-trained deep learning models in object detection and recognition,” *Journal of Information Technology Management*, 14, 4, pp. 40–56, 2022, doi: 10.22059/jitm.2022.88134.
 - [17] A. A. Ojugo and O. Nwankwo, “Spectral-Cluster solution for Credit-Card fraud detection using a genetic algorithm trained modular deep learning neural network,” *JINAV Journal of Information and Visualization*, vol. 2, no. 1, pp. 15–24, Jan. 2021, doi: 10.35877/454ri.jinav274.
 - [18] T. S. Prajwal and I. A. K., “A Comparative Study Of RESNET-Pretrained Models For Computer Vision,” *IC3-2023: Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, pp. 419–425, Aug. 2023, doi: 10.1145/3607947.3608042.
 - [19] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis and O. Chum, “Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5706–5715, doi: 10.1109/CVPR.2018.00598.
 - [20] Y. Li, “The Investigation of DeiT model Based on PaddlePaddle Framework on CIFAR-10 Dataset Image Classification,” in *Advances in computer science research*, 2023, pp. 1062–1067. doi: 10.2991/978-94-6463-300-9_106.
 - [21] R. Khosrowshahli, F. Kheiri, A. A. Bidgoli, H. R. Tizhoosh, M. Makrehchi, and S. Rahnamayan, “Enhancing image retrieval through optimal barcode representation,” *Scientific Reports*, vol. 15, no. 1, Aug. 2025, doi: 10.1038/s41598-025-14576-x.
 - [22] M. Nallappan and R. Velswamy, “Exploring deep learning-based content-based video retrieval with Hierarchical Navigable Small World index and ResNet-50 features for anomaly detection,” *Expert Systems With Applications*, vol. 247, p. 123197, Jan. 2024, doi: 10.1016/j.eswa.2024.123197.
 - [23] M. S. Rao, “Hybrid Deep Learning Approach for Marine Debris Detection in Satellite Imagery Using UNet with ResNext50 Backbone,” *Journal of Applied Science and Technology Trends*, vol. 6, no. 1, pp. 50–60, Jun. 2025, doi: 10.38094/jastt61243.
 - [24] V. S. Anagani, A. Rani, P. Panuganti, and M. Tharangini, “Pancreatic Cancer Detection Using Quaternion Wavelet Transform and Squeeze-and-Excitation Network with SVM Classifier,” *Journal of Applied Science and Technology Trends*, vol. 6, no. 2, pp. 194–202, Aug. 2025, doi: 10.38094/jastt62269.
 - [25] G. Gautam and A. Khanna, “Content Based Image Retrieval System Using CNN based Deep Learning Models,” *Procedia Computer Science*, vol. 235, pp. 3131–3141, Jan. 2024, doi: 10.1016/j.procs.2024.04.296.
 - [26] Z. Chao, S. Cheng, and Y. Li, “Deep internally connected transformer hashing for image retrieval,” *Knowledge-Based Systems*, vol. 279, p. 110953, Sep. 2023, doi: 10.1016/j.knosys.2023.110953.
 - [27] A. P and G. R., “Optimizing visual data retrieval using deep learning driven CBIR for improved human machine interaction,” *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-05478-z.
 - [28] R. Kumar and N. M. M. S., “Enhancing Content-based Image Retrieval Performance through Optimized Feature Selection,” *Engineering Technology & Applied Science Research*, vol. 15, no. 3, pp. 23783–23789, Jun. 2025, doi: 10.48084/etasr.10974.