Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Conv-Guided Mamba Vision Transformer and SPP-DenseNet121 based Hybrid Multi-Feature Fusion and Query Expansion in Content-Based Medical Image Retrieval

Mathana Gopal Arulsamy
Kalasalingam Academy of Research and Education
Pallikonda Rajasekaran Murugan
Kalasalingam Academy of Research and Education
Md. Jakir Hossen
jakir.hossen@mmu.edu.my

Multimedia University
Wai Kit Wong
Multimedia University
Poh Kiat Ng
Multimedia University
Gomathy Nayagam M
Ramco Institute of Technology

Article

Keywords: CBMIR, Endoscopic Bladder Tissue Retrieval, CG-MViT, Multi-Scale Feature Extraction, Query Expansion, Pseudo-Relevance Feedback, Bladder Cancer Imaging, Medical Image Analysis

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10312470/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Conv-Guided Mamba Vision Transformer and SPP-DenseNet121 based Hybrid Multi-Feature Fusion and Query Expansion in Content-Based Medical Image Retrieval

[1]Mathana Gopal Arulsamy, [2*]Pallikonda Rajasekaran Murugan, [3*]Md. Jakir Hossen, [4]Wai Kit Wong, [5]Poh Kiat Ng, [6]M Gomathy Nayagam

[1, 2] Department of Electronics and Communication Engineering,
Kalasalingam Academy of Research and Education, Tamil Nadu, India.

[3,4,5]Faculty of Engineering and Technology,
Multimedia University, Melaka, Malaysia.

[6]Department of Computer Science and Business Systems,
Ramco Institute of Technology, Tamil Nadu, India

[1]m.mathanagopal@klu.ac.in, [2]m.p.raja@klu.ac.in, [3]jakir.hossen@mmu.edu.my
[4]wkwong@mmu.edu.my, [5]pkng@mmu.edu.my, [6]gomathynayagamm@gmail.com

*Corresponding Author: Pallikonda Rajasekaran Murugan (*m.p.raja@klu.ac.in*), Md. Jakir Hossen (*jakir.hossen@mmu.edu.my*)

Abstract: - Content-Based Medical Image Retrieval (CBMIR) is recognized as one of the computer-aided methods in image retrieval systems which can fetch clinically useful images from huge repositories of medical image databases. Endoscopic bladder tissue images have always been difficult to extract from massive image repositories because of the variations within the same class, similarities between different classes, and diverse appearances of the tissues in different imaging modalities. Therefore, this research proposed an innovative novel Multi-Feature Fusion and Query Expansion-Based CBMIR framework using Conv-Guided Mamba Vision Transformer (CG-MViT). At first, input images are processed by image resizing and adaptive contrast enhancement for the sake of better visualization. Then, dual branch feature extraction approach is applied in which SPP-DenseNet121 extracts multi-scale local discriminative features whereas CG-MViT learns long-distance dependencies. The features obtained from both branches are then combined by applying Discriminative Multiple Canonical Correlation Analysis (DMCCA) to obtain discriminative hybrid feature vectors that can be saved in feature database. In case of retrieval, similar processing steps are followed and matched on the basis of composite distance metric. To bridge the gap further and enhance the relevance of the results, a pseudo-relevance feedback-based query expansion approach is employed, and a final re-ranking step is performed. The performance of the proposed framework is tested using two medical image datasets publicly available: the gastrointestinal endoscopic dataset, Kvasir; and the Endoscopic Bladder Tissue dataset. Experiments conducted on the Kvasir database have shown the superiority of the framework, attaining mAP accuracy of 97.72%, 96.81%, 95.78%, 93.15%, and 91.02% with respective retrieval depths of P@5, P@10, P@20, P@50, and P@100. In a similar fashion, the proposed approach has attained mAP accuracies of 92.60%, 90.68%, 84.67%, 70.79%, and 66.82% on the Endoscopic Bladder Tissue database, respectively, with retrieval depths of P@5, P@10, P@20, P@50, and P@100. Additionally, the proposed model has been validated through qualitative assessment and ablation analysis to verify the efficiency of the hybrid features extraction, fusion, and query refinement methods. The experimental results prove that the proposed framework has high robustness and generalizability for clinical image retrieval.

Keywords: CBMIR; Endoscopic Bladder Tissue Retrieval; CG-MViT; Multi-Scale Feature Extraction; Query Expansion; Pseudo-Relevance Feedback; Bladder Cancer Imaging; Medical Image Analysis.

1. INTRODUCTION

The fast expansion in digital health care and medical imaging technology has led to an immense amount of data being generated in the form of medical images [1]. Among the different types of medical imaging, endoscopic imaging is one of the most important in the detection and monitoring of various diseases like gastrointestinal diseases, bladder cancer, and various other diseases related to internal

organs [2,3]. This rise in the amount of data available in medical image databases has increased the need for image retrieval methods. CBMIR is a new domain of research which uses the image contents to search for images similar to one another visually or semantically rather than basing the retrieval process only on textual information [4,5]. The application of CBMIR techniques can help medical practitioners in making diagnoses and treatment plans as well as in medical education by retrieving similar previous cases [6].

The traditional methods of CBMIR focused on manual feature extraction approaches such as color-based description, texture-based description, shape features, Local Binary Patterns (LBP), Histograms of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT) [7-10]. Despite delivering satisfactory results on simpler data, these methods were insufficient for representing complex structures found in medical images. With regard to addressing the shortcomings associated with handcrafted features, several deep learning-based models have gained prominence in the domain of medical image retrieval. CNNs have shown great potential when it comes to learning a hierarchy of representations from massive volumes of data automatically [11]. A number of investigations have used deep learning architectures including VGGNet, ResNet, DenseNet, Inception, and EfficientNet for medical image retrieval tasks by learning discriminative features using CNNs [12, 13]. CNN-based approaches are highly superior compared to handcrafted features due to their powerful ability to learn low and high-level visual representations. Nonetheless, conventional CNN models tend to emphasize learning based on local receptive fields and failed to address the long-term dependency issue associated with medical images [14, 15].

In recent times, transformers have received significant attention in the field of medical image analysis because of their effectiveness in representing global contextual information using self-attention models [16]. ViT-based image retrieval models have been found to perform better when it comes to representing long-distance spatial correlations within images as well as the inter-dependencies among regions within an image [17]. Although transformer-based models are powerful in several ways, they need heavy computational power and huge amounts of data for training purposes to realize their maximum potential [18]. In addition, pure transformer-based techniques may lack the ability to represent local structural information that is required to distinguish different pathological variations.

The other significant problem related to CBMIR is that of integrating heterogeneous feature representations. Many methods have been developed to integrate features derived using various networks in the context of feature extraction. Methods such as concatenation of features, weighted averaging, attention-based feature fusion, and dimensionality reduction are commonly used [19]. These methods lead to redundancy in features and inability to capture any relationship between different sets of features, thus limiting their discrimination power. Other than feature representation difficulties, semantic gap continues to be one of the biggest barriers in CBMIR systems [20]. Methods that use similarity matching based on distances often tend to select similar looking images that are not relevant from the perspective of clinicians [21]. To cope with this problem, some studies have tried to apply relevance feedback, query expansion, and re-ranking algorithms in their search models [22, 23]. Even though these approaches enhance retrieval performance, they still have shortcomings in that they involve manual work or lack utilization of semantics of retrieved images.

In light of these problems, the objective of this study is to propose a framework for CBMIR which uses complementary feature representation, discriminative feature fusion, and query refinement approaches to retrieve bladder tissues from endoscopic images with high accuracy. In this paper, we seek to use efficient feature learning techniques and retrieval optimization techniques in our proposed framework to overcome the semantic gap and discriminable features to obtain the most relevant retrieval results.

1.1 Major Contributions

The major contributions of the proposed research are summarized as follows:

- A new CBMIR paradigm for retrieving endoscopic bladder tissue images has been introduced by allowing clinicians to access useful images from large medical databases and facilitating computer-aided decision making.
- Two-stream feature extraction networks are introduced that combine SPP-DenseNet121 and CG-MViT to achieve a comprehensive representation of the image.

This includes both local features at different scales as well as long-term contextual information.

- SPP is utilized in DenseNet121 to improve the process of extracting scale-invariant features and represent abnormalities within bladder tissue of different scales and structures.
- A CG-MViT block is used to learn global semantic relations and long-range spatial relations with less computation as compared to transformer-based models.
- The DMCCA-based fusion method is proposed for heterogeneous feature representations to obtain discriminative and redundant-free feature space.
- The PRF-based query expansion approach is utilized for improving the query representation and solving the ambiguity problem and semantic gap between visual and clinical aspects of the images.
- A thorough ablative study has confirmed the importance of each framework constituent by validating the contribution of CG-MViT, DMCCA fusion, and query expansion using PRF towards improving retrieval efficiency.

2. RELATED WORK

CBMIR has gained substantial attention in the past few years owing to its capabilities to help clinicians in diagnosis and decision-making by visually similar medical images in extensive databases. Different methods have been used in the past to address feature extraction, feature fusion, similarity computation, and query refinement issues for better retrieval performance. Traditional methods focused more on manual feature extraction using various texture and shape descriptors, while the recent studies focus on using deep learning techniques, transfer learning, and foundation models for feature extraction. There are some limitations in current techniques, like the inability to properly model global context information, feature redundancy issues, high computational costs, and a semantic gap between visual features and clinical semantics that limit their performance. Table I provides a summary of prominent CBMIR techniques in terms of their strengths and weaknesses.

Table I Summary of Existing CBMIR Methods and their Limitations

Reference	Methodology	Major Contribution	Merits	Demerits
Badanidiyoor et al. (2026) [24]	Local texture feature-based CBMIR	Investigated the effectiveness of local texture descriptors for medical image retrieval.	Simple implementation, low computational cost, interpretable feature representation.	Limited ability to capture high-level semantic information and complex image structures; retrieval performance decreases for heterogeneous datasets.
Singh and Singh (2025) [25]	Deep feature fusion with dimensionality reduction for gastric image retrieval	Combined deep learning features and dimensionality reduction techniques to improve retrieval efficiency.	Enhanced discriminative feature representation and reduced feature dimensionality.	May lose important discriminative information during dimensionality reduction; limited global context modelling.
Humadi et al. (2025) [26]	Adaptive two-tier deep learning with dynamic similarity fusion	Proposed a multi-stage retrieval framework with adaptive similarity fusion and classification.	Improved retrieval robustness through dynamic similarity computation.	Increased computational complexity and dependence on multiple processing stages.
Kim et al. (2025) [27]	Foundation model-based retrieval for gastrointestinal endoscopy	Utilized foundation models to improve retrieval performance in endoscopic image datasets.	Strong semantic representation capability and improved retrieval accuracy.	Requires extensive computational resources and large-scale training data; limited deployment efficiency.

Reference	Methodology	Major Contribution	Merits	Demerits
Ahmed and Malebary (2020) [28]	Query expansion using top-ranked images	Introduced query expansion based on initial retrieval results to improve retrieval relevance.	Effectively reduces retrieval ambiguity and improves semantic consistency.	Retrieval performance strongly depends on the quality of initial retrieval results.
Rani et al. (2026) [29]	Optimized intermediate layer feature-based CBMIR	Exploited intermediate deep network representations for enhanced retrieval performance.	Captures richer hierarchical information than final-layer features.	Intermediate features may contain redundant information and require additional optimization.
Ahmed et al. (2022) [30]	Pre-trained CNN-based CBMIR	Applied transfer learning using pre-trained CNN models for medical image retrieval.	Reduced training requirements and improved feature extraction capability.	CNN features mainly focus on local patterns and may overlook long-range dependencies.
Alzahrani et al. (2024) [31]	Multiple pre-trained CNN feature extraction models	Combined features from multiple CNN architectures for improved retrieval accuracy.	Utilizes complementary information from different deep models.	High feature dimensionality and increased computational burden due to multiple networks.
Ahmed et al. (2023) [32]	Relative difference-based similarity measure	Proposed an improved similarity measurement strategy for CBMIR systems.	More effective similarity assessment compared to conventional distance measures.	Retrieval performance remains dependent on the quality of extracted features and feature representations.

2.1 Discussion and Research Gap

The survey on the literature reveals that recent CBMIR systems have increasingly moved away from manual engineering of feature descriptors and toward deep learning-based feature extraction and feature fusion techniques. Even though CNN-based models have shown improved results compared to conventional systems, they still fail to capture long-range contextual relations and semantic correlations between medical images. Although foundation models and transformer-based models are able to enhance global feature learning, they often come at a cost of heavy computations and extensive training datasets. Moreover, many of the available feature fusion models are prone to redundancy and fail to leverage all the information in heterogenous feature spaces. Another common problem in current CBMIR systems is that of the semantic gap, leading to retrieval of images which are similar in appearance but do not share much clinical relevance.

3. PROPOSED METHODOLOGY

The overall architectural design of the CBMIR model can be broadly divided into two main steps, offline feature learning and online image search and it is illustrated in Figure 1. In the offline step, all database images will initially go through the preprocessing step, whereby resizing and contrast enhancement techniques are used to standardize and enhance the quality of the images. The enhanced images will then go through feature extraction using two different complementary networks. The SPP-DenseNet121 branch can capture the multi-scale hierarchically structured features that are capable of representing local anatomy information, texture patterns, and even lesions in the images. On the other hand, the Conv-Guided Mamba Vision Transformer (CG-MViT) branch can represent the global contextual information and spatial dependencies between different areas in the endoscopic images. Due to the complementary nature of the two feature sets, DMCCA is employed to fuse these two sets of features through maximization of the correlations between them. The hybrid feature vectors generated

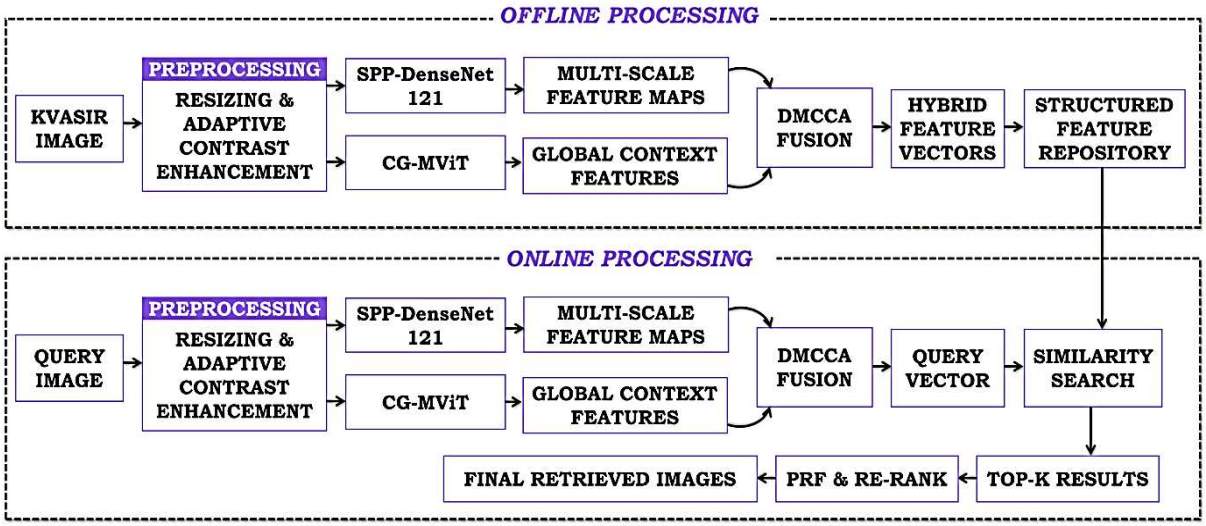


Fig. 1. Overall architectural design of the CBMIR model

from these processes are then placed in a feature repository, which acts as the searchable database for the retrieval application. During the online phase, the new query image is subjected to the same pre-processing and dual branch feature extraction process to produce its own query feature vector. The similarity between the feature vectors in the database and the query image vector is evaluated using a combined similarity measure method. Using these similarity measurements, relevant images are initially retrieved. Afterwards, Pseudo-Relevance Feedback (PRF) technique is employed to improve and enhance the query model by analyzing the top-ranked retrieval outcomes, thus improving query ambiguity reduction and minimizing the gap that exists between low-level image visual features and high-level medical concepts. The images obtained from the process are then ranked based on the improved similarity measure score values to generate the final set of relevant medical images. With the use of multi-scale convolutional learning, global context modelling, discriminative feature fusion, similarity measure, and query expansion, the proposed framework offers an effective CBMIR solution.

3.1 Offline Feature Learning and Hybrid Feature Repository Construction

The initial stage of the proposed CBMIR approach is the offline processing stage as shown in Figure 2 deals with building the feature bank based on the medical image database. This stage has to be executed only one time when the system is trained/ indexed to create stable feature vectors that include both local and global features of the gastrointestinal endoscopy images.

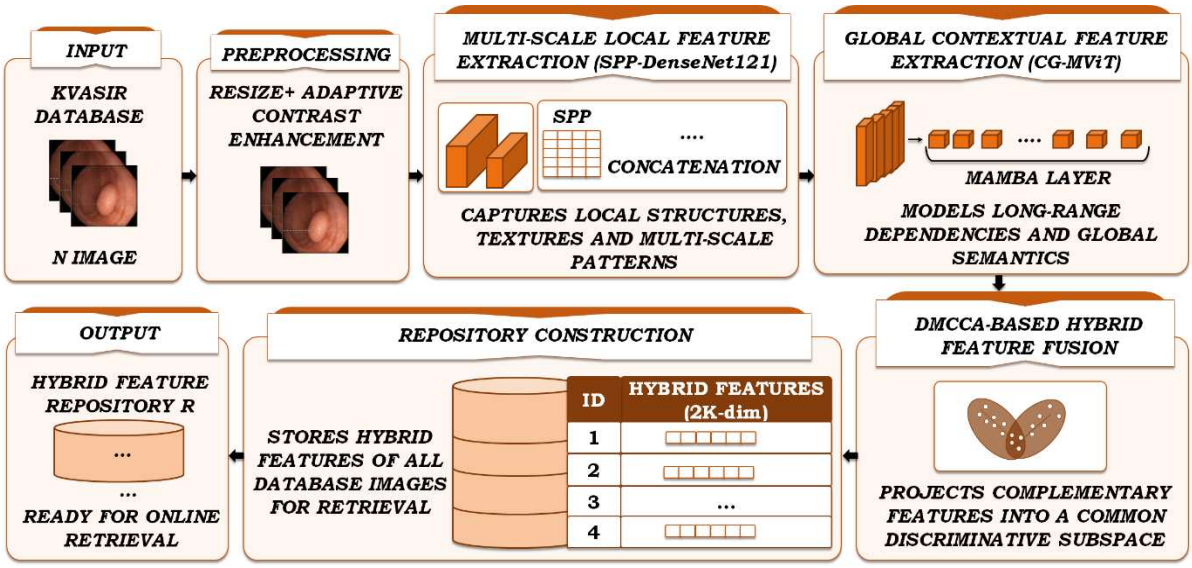


Fig. 2. Architecture of offline learning stage

3.1.1 Image Preprocessing and Contrast Enhancement

Consider the Kvasir image repository defined by $\mathcal{D} = \{I_1, I_2, \dots, I_N\}$. Here, the total number of images in the database is termed as N and $I_i$represents the i^{th}endoscopic image. Due to the presence of different illumination effects, low contrast levels, and scaling factors in medical images, the dataset images are resized to the spatial resolution $H \times W$ and contrast enhancement. The mathematical expression for the preprocessing step is given below:

$$I_i^p = \Phi(I_i) \tag{1}$$

Here, the pre-processed image is stated as I_i^p and the combined resizing and adaptive contrast enhancement function is defined as $\Phi(\cdot)$. The intensity transformation in adaptive contrast enhancement is described below:

$$I_i^e(x,y) = \frac{I_i(x,y) - I_{\min}}{I_{\max} - I_{\min}} \tag{2}$$

Here, the original and enhanced pixel intensity is stated as $I_i(x,y)$ and $I_i^e(x,y)$ respectively.

3.1.2 Multi-Scale Local Feature Extraction Using SPP-DenseNet121

The feature extraction approach should be able to capture both the detailed and coarse features within a structure. In order to solve this problem, the SPP-DenseNet121 network is used as the first feature extractor in the framework. DenseNet121 has been chosen due to the dense connection strategy implemented by DenseNet121, which allows feature transfer directly between any two preceding and succeeding layers. While conventional neural networks have information flow from one layer to another in a sequential order, the DenseNet model allows concatenation of earlier layer outputs with their usage across the entire architecture. The dense connection strategy solves the problem of vanishing gradients, makes features reusable, and allows for retention of high and low-level image features. The output from the l^{th} layer in a DenseNet with L layers can be expressed as follows:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \tag{3}$$

Here, the output feature map of the l^{th}layer is stated as x_l, composite transformation function is stated as $H_l(\cdot)$ and the feature concatenation from all preceding layers is defined as $[x_0, x_1, \dots, x_{l-1}]$. After the input image passes through the DenseNet121 backbone, the resultant feature map is generated, and it is defined below:

$$F_D = \Psi_D(I_i^p) \tag{4}$$

Here, the DenseNet121 feature extraction function is stated as $\Psi_D(\cdot)$, the resulting feature tensor is defined as F_D. Despite the high discriminative power of DenseNet121 for feature extraction, similarity between the images after the retrieval process impacted by differences in the size of lesions. Since in an actual endoscopic procedure, lesions can appear in small or large regions within the image frame, the conventional global pooling technique falls short of retaining such multi-scale information. To solve this issue, the Spatial Pyramid Pooling (SPP) layer has been used. The SPP layer divides the input feature maps into several spatial bins with different scales and performs an independent pooling operation in each bin. Assuming that l denotes the pyramid level, the pooling output at pyramid level $l$is is obtained as follows:

$$P_l = SPP_l(F_D) \tag{5}$$

The pooled feature representation is stated as P_l. For instance, the feature map may be subdivided into grids of sizes 1×1, 2×2, and 4×4. The 1×1 size grid is responsible for representing global image details, whereas the other two sizes help preserve the local lesion features. The multi-level representation is generated using feature concatenation as follows:

$$F_{MS} = [P_1, P_2, \dots, P_L] \tag{6}$$

Here, the total number of pyramid levels is stated as L and the multi-scale DenseNet feature vector is defined as F_{MS}.

Figure 3 shows the multi-scale local feature extraction branch of the proposed framework. Firstly, the processed endoscopic image is passed as an input to the DenseNet121 base model that is made up of several dense blocks connected to each other via transition layers. With the help of dense connectivity, the network allows each layer to receive feature representation from all previous layers,

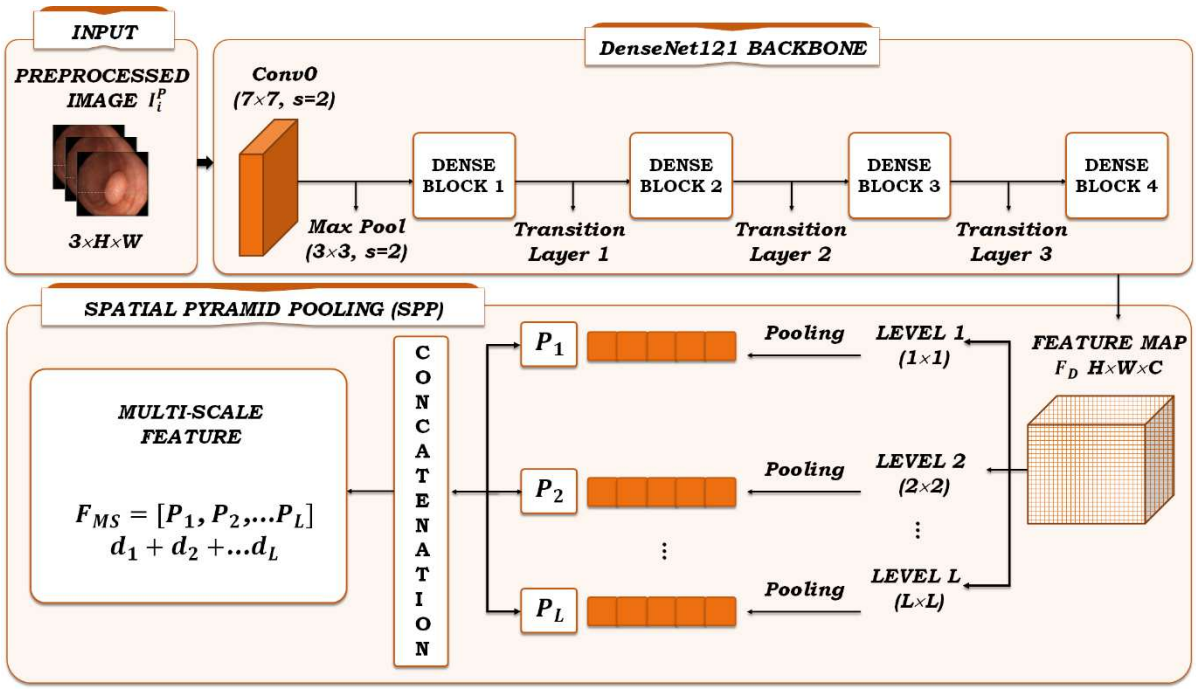


Fig. 3. Multi-Scale Local Feature Extraction Using SPP-DenseNet121

thus leading to efficient reuse of features and maintenance of both low-level and high-level visual information. By processing the image through dense blocks, discriminative feature representation regarding lesion margins, mucosa texture, vessels, and anatomy is learned, leading to a deep convolution feature map.

To obtain the different scales of abnormalities, the resultant feature map is then passed through the SPP layer. Spatial Pyramid Pooling works by dividing the feature map into several spatial partitions at different levels of the pyramid structure and conducting pooling separately for each spatial partition. This approach allows us to obtain both global and local features at different scales. The pooled feature maps are then combined together to produce a multi-scale feature vector representation in an effective manner.

3.1.3 Global Contextual Feature Extraction Using Conv-Guided Mamba Vision Transformer (CG-MViT)

In medical image retrieval, however, diagnostic interpretation frequently relies on understanding relationships among multiple image regions rather than isolated local patterns. To overcome this limitation, the proposed framework incorporates a CG-MViT as the second feature extraction branch.

The main aim of CG-MViT model is to generate contextual information at the global level, which serves as an addition to the local structural information obtained from SPP-DenseNet121. The vision transformers traditionally make use of self-attention mechanisms, which are quite expensive for computation.

However, the model Mamba uses selective state space approach to capture the long-range dependencies with linear computational complexity. Firstly, the pre-processed input is fed into a light-weight convolutional embedding module:

$$F_C = \Psi_C(I_i^p) \tag{7}$$

The convolution-guided feature map is stated as F_C and the convolutional embedding function is defined as $\Psi_C(\cdot)$. There are two main roles played by the convolution-based embedding layer. The first one is that it retains local spatial consistency which could have been disrupted during tokenization. Second, the layer gives inductive bias for the learning of anatomy and lesions boundaries. As a result, local information remains clinically significant as inputs for the Mamba framework. The resulting feature map is converted into a token sequence, and it is defined below:

$$T = \{t_1, t_2, \dots, t_n\} \tag{8}$$

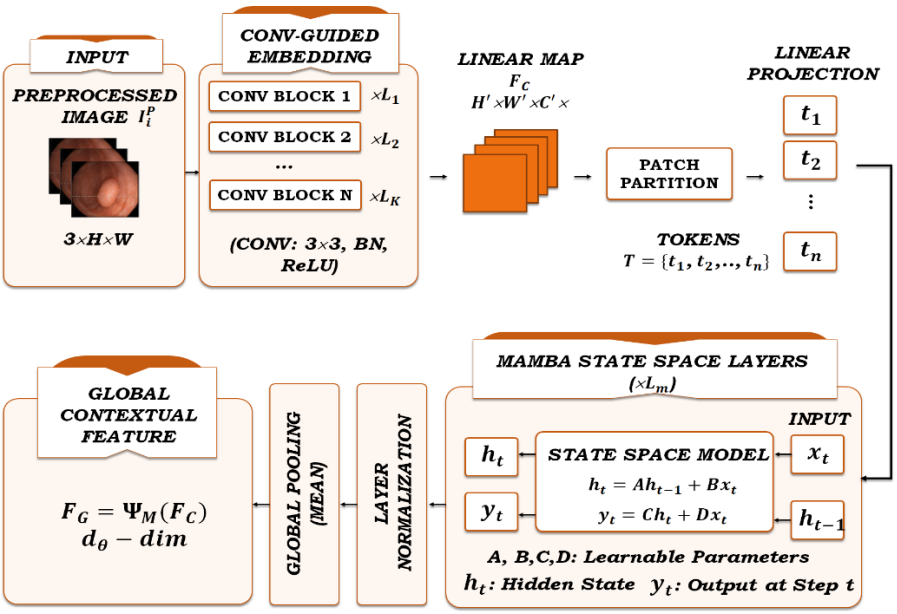


Fig. 4. Global Contextual Feature Extraction Using Conv-Guided Mamba Vision Transformer (CG-MViT)

Here, the set of tokens, number of tokens are specified as T and n respectively. This token sequence is then processed by the Mamba State Space Model, and the hidden state update equation is defined as follows:

$$h_t = Ah_{t-1} + Bx_t \tag{9}$$

The hidden state at position t and the previous hidden state is termed as h_t and h_{t-1} respectively. Also, the current input token is defined as x_t and the learnable state transition matrices are defined as $A$and B consecutively. Therefore, the output representation of this process is given below:

$$y_t = Ch_t + Dx_t \tag{10}$$

The learnable projection matrices are defined as $C$and D. Following several Mamba layer applications, the output of the final contextual representation is defined below:

$$F_G = \Psi_M(F_C) \tag{11}$$

Here, the Mamba feature learning operation is defined as $\Psi_M(\cdot)$ and the global contextual feature vector is defined as F_G.

Figure 4 depicts the architecture of global contextual feature extraction based on the proposed CG-MViT model. After passing the pre-processed image through multiple convolutional embedding layers that extract local structure information while maintaining spatial connectivity, the obtained convolution guided features are mapped into a sequence of visual tokens via patching and linear projections. The output visual tokens are fed into the Mamba State Space layers, which perform modeling of long-term correlations and global context among image patches. As opposed to the traditional architectures of transformers which depend on the self-attention process, the Mamba architecture uses selective state-space model to effectively pass information between spatial coordinates. The state updates allow contextual information came from the distant areas of images to be utilized for building representations of each token. After layer normalization and global pooling, the global contextual feature vector is obtained which captures semantic information and helps to complete the multi-scale features captured by the SPP-DenseNet121 branch.

The integration of SPP-DenseNet121 and CG-MViT allows the proposed framework to utilize all the local lesion features, multiscale structural features, and global contextual semantic features simultaneously. The proposed approach for feature learning acts as the starting point for the next phase, which involves integrating the two feature spaces using DMCCA.

3.1.4 DMCCA-Based Hybrid Feature Fusion

The feature representations obtained from the SPP-DenseNet121 and CG-MViT branches carry complimentary knowledge about the visual features of endoscopic imaging for gastrointestinal tracts.

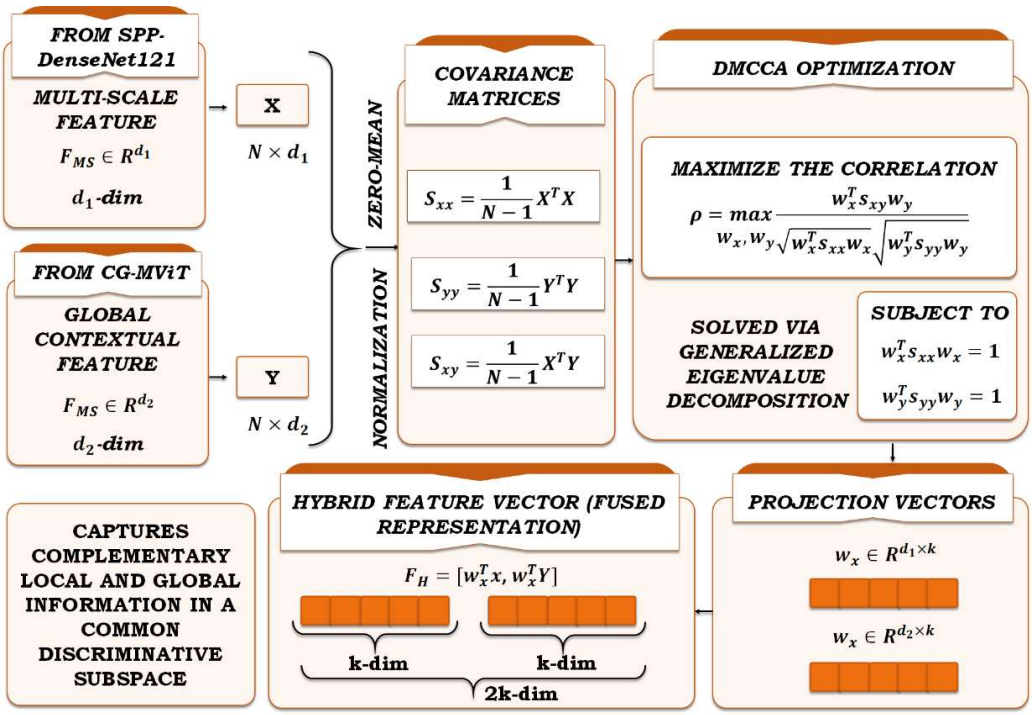


Fig. 5. DMCCA based hybrid feature fusion

The use of DMCCA which overcomes the limitations of canonical correlation analysis (CCA) based feature fusion [33], provides the projection of different feature spaces into a discriminative subspace. This results in a unified feature representation capturing both local and global information. Let, $X = F_{MS}$ and $Y = F_G$ define the multi-scale feature vector learned using the SPP-DenseNet121 model and global context feature vector obtained using the CG-MViT architecture, respectively. The DMCCA objective function is formulated as

$$\rho = \max_{w_x, w_y} \frac{w_x^T S_{xy} w_y}{\sqrt{w_x^T S_{xx} w_x} \sqrt{w_y^T S_{yy} w_y}} \tag{12}$$

Here, the canonical correlation coefficient is defined as ρ. The covariance matrix of feature set X and Y is defined as S_{xx} and S_{yy}. The cross-covariance matrix between $X$and Y is defined as S_{xy} and the projection vectors associated with the two feature spaces is defined as $w_x$and w_y consecutively. Through this optimization procedure, DMCCA is able to discover high correlations and discriminative components in both representations and project them into a common latent space. As a result, the fused feature representation turns out to be concise and informative with less redundant features. The final hybrid feature vector is obtained as follows:

$$F_H = [w_x^T X, w_y^T Y] \tag{13}$$

Here, the hybrid feature representation is specified as F_H. The projected multi-scale feature component and the global contextual feature component are defined as $w_x^T X$ and $w_y^T Y$. Consequently, the representation obtained by this method is able to provide a better understanding of the medical image contents and thus improves retrieval performance. Therefore, the hybrid feature fusion process is illustrated in Figure 5.

3.1.5 Hybrid Feature Repository Construction

The resulting hybrid feature vectors after DMCCA fusion are then stored in a structured storage space, which acts as the searchable knowledge source for the retrieval system. The goal of this step is to maintain the feature vectors of all images from the database so that feature extraction does not have to be done again and again in retrieval phase. After that, all vectors will be stored within the repository along with corresponding images' metadata and IDs. The repository structure may be expressed as follows:

$$\mathcal{R} = \{(F_H^{(1)}, L_1), (F_H^{(2)}, L_2), \ldots, (F_H^{(N)}, L_N)\} \tag{14}$$

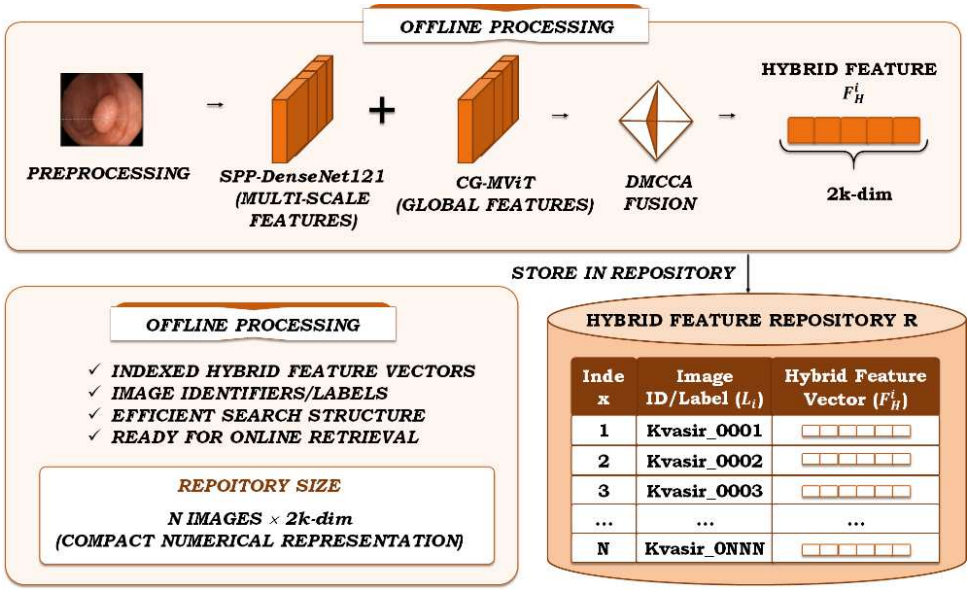


Fig. 6. Hybrid feature repository construction

Here, the hybrid feature repository is specified as $\mathcal{R}$, the total number of database images is denoted as N, hybrid feature vector of the i^{th}image is specified as $F_H^{(i)}$ and the corresponding image label, identifier, or metadata is specified as L_i. The repository construction is illustrated in Figure 6.

3.2 Online Query Processing and Retrieval

Online retrieval phase is concerned with the extraction and selection of those medical images that are most relevant from the existing hybrid feature repository. As opposed to the offline phase which deals with feature representation and storage of database images, the online phase involves processing of a new image queried by the user. In this case, the aim is to convert this query image into a discriminative feature representation by comparing it with those already represented in the existing repository of features in order to retrieve visually and semantically similar images. To facilitate effective retrieval and minimize semantic gap between image features and medical concepts, the proposed framework employs similarity-based matching, pseudo relevance feedback, query expansion and re-ranking techniques.

3.2.1 Query Feature Generation

Once a search image is entered into the retrieval system, the same preprocessing steps carried out in the off-line phase are performed on it. The uniformity of preprocessing processes guarantees that features extracted from database images and query images are consistent. Let the query image be represented as I_q. The preprocessing operation is expressed as follows:

$$I_q^p = \Phi(I_q) \tag{15}$$

The pre-processed query image is stated as I_q^p and the resizing and adaptive contrast enhancement is defined as $\Phi(\cdot)$.The pre-processed query image is then fed into both feature extraction paths. The SPP-DenseNet121 branch extracts the following multi-level local features:

$$F_{MS}^q = \Psi_D(I_q^p) \tag{16}$$

Here, the query multi-scale feature vector is termed as F_{MS}^q . Likewise, the CG-MViT channel extracts global context-aware features and it is defined below:

$$F_G^q = \Psi_M(I_q^p) \tag{17}$$

The query global contextual feature vector is stated as F_G^q. Now, these hybrid feature vectorsF_H^q are combined using the DMCCA projections learned in the offline phase, and it is stated below:

$$F_H^q = [w_x^T F_{MS}^q,\ w_y^T F_G^q] \tag{18}$$

The combined feature vector from the two branches becomes the retrieval descriptor for matching.

3.2.2 Similarity Computation

Once the vector representation for the query has been created, the similarity of this vector is calculated with all images present in the database. This is because Euclidean distance calculates the difference in the absolute values of features, while the cosine similarity determines the direction of two feature vectors. Let the query feature vector and the feature vector of the i^{th}database image are stated as F_H^q and $F_H^{(i)}$ respectively. Therefore, the Euclidean distance is evaluated using the below expression:

$$D_E(i) = \sqrt{\sum_{j=1}^{d}(F_{H,j}^q - F_{H,j}^{(i)})^2} \tag{19}$$

Then, the cosine similarity is computed using below formulation:

$$S_C(i) = \frac{F_H^q \cdot F_H^{(i)}}{\|F_H^q\| \|F_H^{(i)}\|} \tag{20}$$

Here, the vector magnitude is stated as $\|\cdot\|$ and the final composite similarity score is examined using below expression:

$$S(i) = \alpha S_C(i) + (1-\alpha)\left(\frac{1}{1+D_E(i)}\right) \tag{21}$$

The final similarity score is stated as $S(i)$ and the weighting coefficient is denoted as α. If the value of S(i) is larger, the image is more similar to the query image.

3.2.3 Initial Retrieval

All the images in the database are sorted in terms of their similarities in decreasing order. The ranking function can be represented as follows:

$$Rank = Sort(S(i)) \tag{22}$$

The descending sorting of similarity scores is defined as $Sort(\cdot)$ and the top-K most similar images are selected and it is defined below:

$$R_K = \{I_1, I_2, \dots, I_K\} \tag{23}$$

Here, the number of retrieved images is defined as K . The images that have been found during the initial retrieval process are considered as candidate relevant images, which need further processing. Even though the initial image search retrieves images that are visually similar to each other, irrelevant images might also be present among the retrieved images, since visual similarity does not necessarily imply semantic similarity. Therefore, an additional refinement stage is introduced.

3.2.4 Pseudo-Relevance Feedback (PRF)

Pseudo-Relevance Feedback (PRF) is used for reducing retrieval ambiguity, as well as filling the semantic gap between visual content and medical terms. Unlike traditional relevance feedback mechanisms which rely on user interaction, PRF automatically considers that the most highly ranked search results are relevant to the search query. Let $R_P = \{I_1, I_2, \dots, I_P\}$ denote the top-P retrieved images. The average feature representation of these images can be evaluated using below expression:

$$F_{PRF} = \frac{1}{P}\sum_{i=1}^{P} F_H^{(i)} \tag{24}$$

Here, the feedback feature vector is defined as F_{PRF}. This vector reflects the most dominant features of the highly ranked retrieved data and can be considered an estimation of the query's semantic intention.

3.2.5 Query Expansion

After receiving the feedback vector, the original feature vector for the query is augmented to include the information of consensus obtained from the top ranked images. Augmented feature vector is expressed below:

$$F_{QE} = (1-\beta)F_H^q + \beta F_{PRF} \tag{25}$$

Here, the expanded query feature vector is defined as F_{QE} and β denotes the influence of feedback information. With the augmented query vector, the retrieved target becomes more realistically described and robust against similarity calculation.

3.2.6 Final Re-Ranking

Once more, the augmented query vector is compared against all the feature vectors of the database using the composite similarity measure. The revised similarity score is then calculated using below expression:

$$S_{new}(i) = \alpha S_C(F_{QE}, F_H^{(i)}) + (1-\alpha)\left(\frac{1}{1+D_E(F_{QE}, F_H^{(i)})}\right) \tag{26}$$

Here, the refined similarity score is stated as $S_{new}(i)$ and the images stored in the database are then re-ranked based on their new scores:

$$Rank_{final} = Sort(S_{new}(i)) \tag{27}$$

Therefore, the final retrieval result set is expressed as follows:

$$T = \{R_1, R_2, \ldots, R_K\} \tag{28}$$

Here, the final ranked retrieval output is stated as T and the k^{th}most relevant retrieved image is stated as R_k respectively. Due to the combination of query feature generation, hybrid similarity measurement, pseudo-relevance feedback, query expansion, and re-ranking, the online retrieval process is greatly improved through enhanced precision and semantic relevance. Through continual refinement of the generated query feature by using information extracted from the initial retrieval of images, the system avoids any ambiguities in retrieval and resulting in highly precise retrieval that matches well with the content of the query image.

The online query processing and retrieval phase is illustrated in Figure 7. It shows the process for the entire sequence of online querying process within the proposed CBMIR framework. In the first place, an input image is processed through preprocessing, which involves resize, adaptive contrast, and normalization processes. In parallel, the two branches of the network, SPP-DenseNet121 and CG-MViT branches, process the pre-processed image to produce local and global feature representations, respectively.

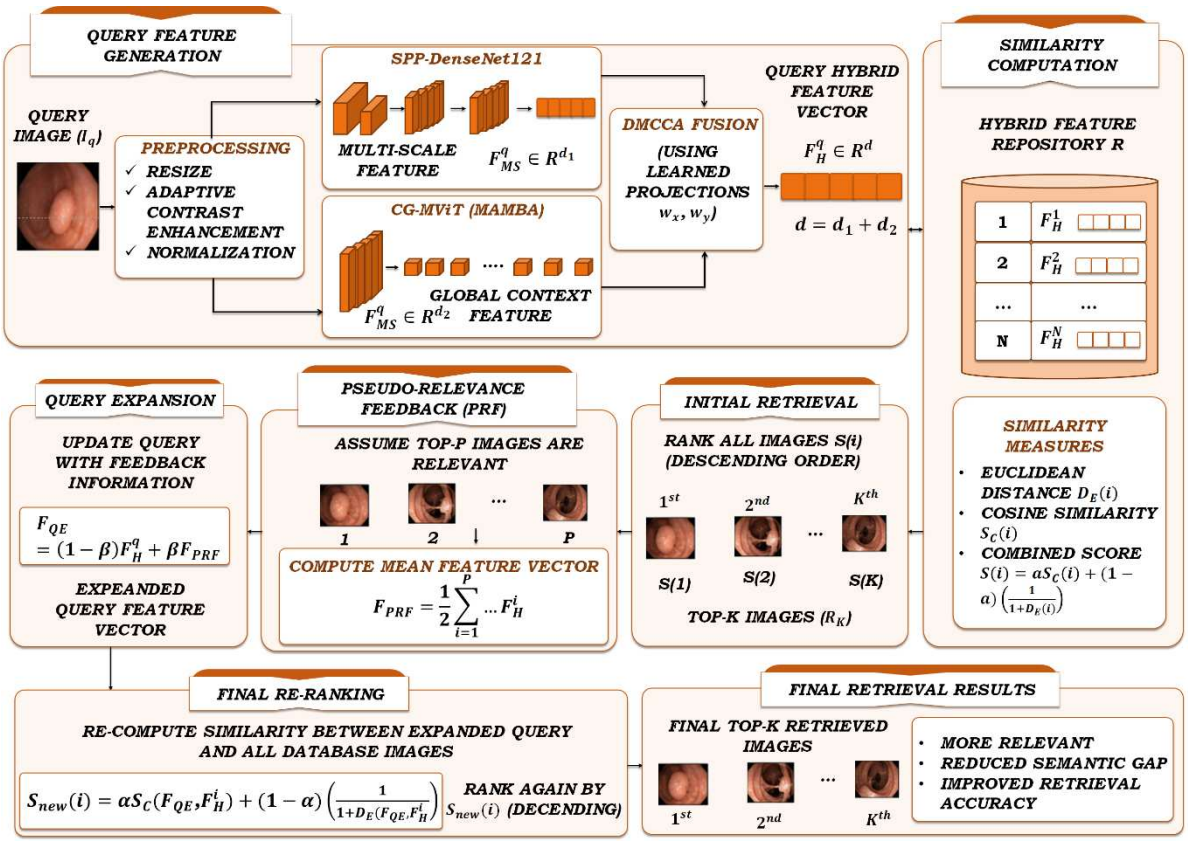


Fig. 7. Online query processing and retrieval

Table II Multi-Feature Fusion and Query Expansion-Based CBMIR Framework

Input:

- Endoscopic image database $D = \{I_1, I_2, \ldots, I_N\}$
- Query image I_q
- Retrieval depth K
- PRF feedback parameter β

Output:

- Top-K retrieved medical images R_K

Offline Repository Construction

1. **For each database image** $I_i \in D$
2. Apply preprocessing:
$$I_i^p = \Phi(I_i)$$
where $\Phi(\cdot)$denotes resizing and adaptive contrast enhancement.
3. Extract multi-scale local features using SPP-DenseNet121:
$$F_{MS}^{(i)} = \Psi_D(I_i^p)$$
4. Extract global contextual features using CG-MViT:
$$F_G^{(i)} = \Psi_M(I_i^p)$$
5. Perform DMCCA-based feature fusion:
$$F_H^{(i)} = [w_x^T F_{MS}^{(i)}, w_y^T F_G^{(i)}]$$
6. Store fused feature vector $F_H^{(i)}$in repository $\mathcal{R}$.
7. **End For**
8. Construct hybrid feature repository:
$$\mathcal{R} = \{F_H^{(1)}, F_H^{(2)}, \ldots, F_H^{(N)}\}$$

Online Query Processing and Retrieval

9. Input query image I_q.
10. Apply preprocessing:
$$I_q^p = \Phi(I_q)$$
11. Extract query multi-scale features:
$$F_{MS}^q = \Psi_D(I_q^p)$$
12. Extract query global contextual features:
$$F_G^q = \Psi_M(I_q^p)$$
13. Generate query hybrid feature vector:
$$F_H^q = [w_x^T F_{MS}^q, w_y^T F_G^q]$$
14. **For each repository feature vector** $F_H^{(i)}$
15. Compute Euclidean distance:
$$D_E(i) = \sqrt{\sum_{j=1}^{d} (F_{H,j}^q - F_{H,j}^{(i)})^2}$$
16. Compute cosine similarity:
$$S_C(i) = \frac{F_H^q \cdot F_H^{(i)}}{\| F_H^q \| \| F_H^{(i)} \|}$$
17. Calculate composite similarity score:
$$S(i) = \alpha S_C(i) + (1 - \alpha)\left(\frac{1}{1 + D_E(i)}\right)$$
18. **End For**

19. Rank all database images according to $S(i)$.
20. Retrieve top-P images:

$$R_P = \{I_1, I_2, \dots, I_P\}$$

Pseudo-Relevance Feedback and Query Expansion

21. Compute feedback feature vector:

$$F_{PRF} = \frac{1}{P}\sum_{i=1}^{P} F_H^{(i)}$$

22. Update query representation:

$$F_{QE} = (1-\beta)F_H^q + \beta F_{PRF}$$

23. Recompute similarity scores between $F_{QE}$and all repository features.
24. Re-rank retrieved images.
25. Select final top-K results:

$$R_K = \{R_1, R_2, \dots, R_K\}$$

26. Return R_K.

The learned DMCCA projections will fuse these two feature representations to obtain a hybrid query representation. Subsequently, the query feature is compared with all hybrid feature vectors in the repository using a combination of Euclidean distance and cosine similarity measures. According to the calculated similarities, a list of initial images is selected. To increase the quality of the retrieval process, a technique referred to as positive relevance feedback (PRF) is assumed. According to the PRF technique, the images at the top of the ranking list are considered relevant, and their mean feature representation is computed. Such information is useful for enhancing the initial feature vector of the query through query expansion. Finally, new similarity scores are calculated, and the ranking list is re-organized. In effect, the proposed method is able to retrieve a more accurate and semantically consistent set of medical images.

4. RESULTS AND DISCUSSION

4.1 Dataset Description (Kvasir)

The suggested medical image retrieval approach will be tested on the Kvasir dataset that is suitable for developing computer systems used in gastrointestinal disease detection. This dataset contains images taken via endoscopy from various sections of the GI tract and provides a broad range of anatomical references and pathological information typical for routine medical examination. Kvasir is especially important to be used in medical imaging studies due to the availability of detailed annotations made by qualified gastroenterologists. The data set consists of eight classes of images, where each normal anatomical structure consists of 500 images with a total of 4,000 images. Such classes are considered important gastrointestinal diseases and anatomical landmarks that can be widely employed in studies related to disease diagnosis, classification, and content-based image retrieval. The sample image of this dataset is illustrated in Figure 8.

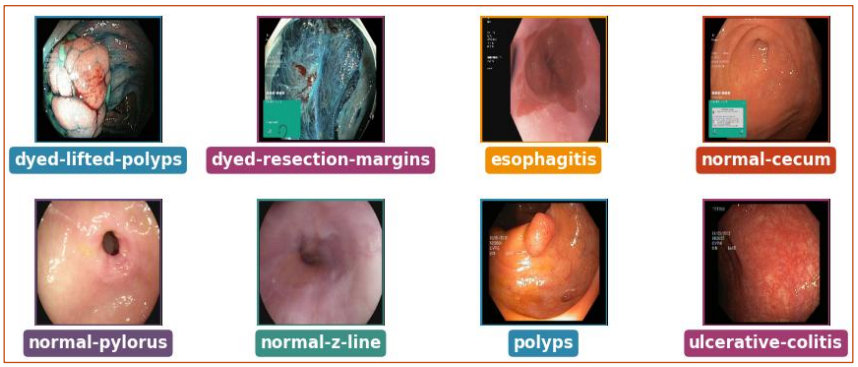


Fig. 8. Sample images of Kvasir dataest

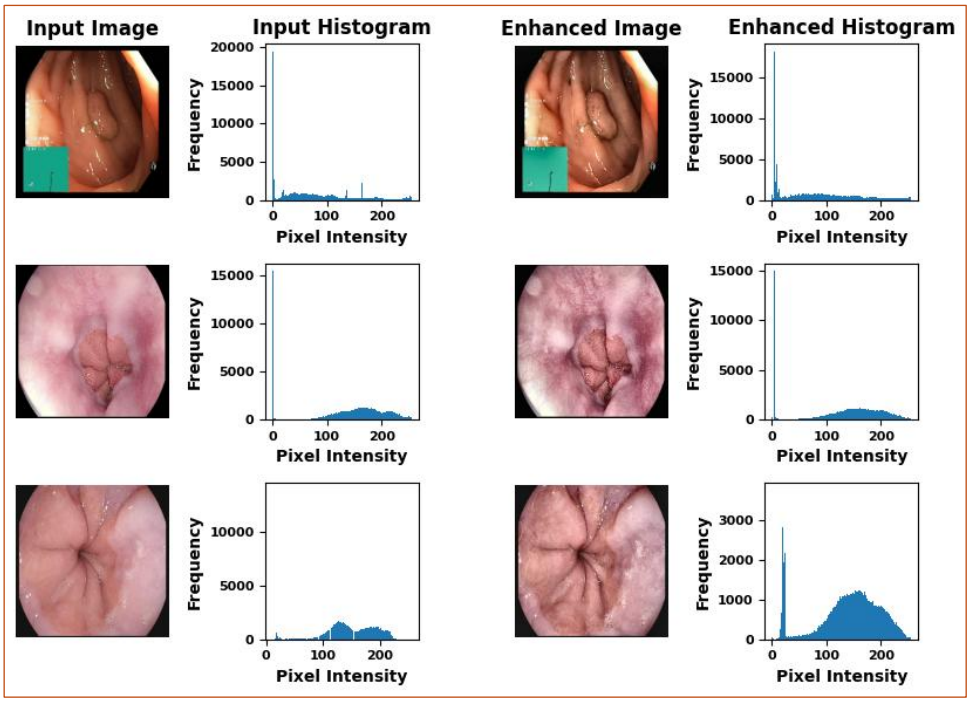


Fig. 9. Visual and Histogram-Based Analysis of Adaptive Contrast Enhancement on Kvasir Endoscopic Images

4.1.1 Effectiveness of Image Preprocessing and Contrast Enhancement

The influence of the preprocessing technique on the sample endoscopic images taken from the Kvasir dataset is depicted in Figure 9. This figure shows the comparison between the input images with their histograms and the resulting images after applying adaptive contrast improvement. It is evident that the preprocessing step has contributed greatly to improving the quality of image in terms of increasing contrast, minimizing inconsistencies in the illumination conditions, and highlighting the relevant structures present in the image.

Moreover, the histogram distribution suggests that there has been an improved distribution of the intensity of pixels throughout the available dynamic range, suggesting improvement in terms of contrast usage and brightness normalization. This is especially helpful in the case of gastrointestinal images, where features become more easily identifiable and allowing for better feature extraction from the image by employing following stages namely the SPP-DenseNet121 module and the CG-MViT module.

4.1.2 Qualitative Analysis of Medical Image Retrieval Performance

Figure 10 shows the qualitative retrieval performance achieved by the proposed CBMIR framework with regard to four different categories of queries selected from the Kvasir dataset. The top 10 similar images retrieved by the framework from the database are shown for each query image based on their similarity score. One can see that most of the retrieved images correspond to the same category as the query image and it indicates the successfulness of the proposed hybrid representation model in identifying local and global features of the lesions.

For the category “Normal-Pylorus,” all top-retrieved images are highly consistent in terms of their structural and visual resemblance. The same can be said for the “Ulcerative-Colitis” category, in which the retrieved images have consistent patterns of inflammation, mucosal injury, and discoloration. The retrieved images under the categories “Polyps” and “Dyed-Lifted-Polyps” also have high consistency in the preservation of structural properties that characterize the lesion. All this shows that the use of SPP-DenseNet121, CG-MViT, DMCCA for feature-level fusion, and Pseudo Relevance Feedback allows for distinguishing between various diseases of the gastrointestinal tract and performing accurate image retrieval.

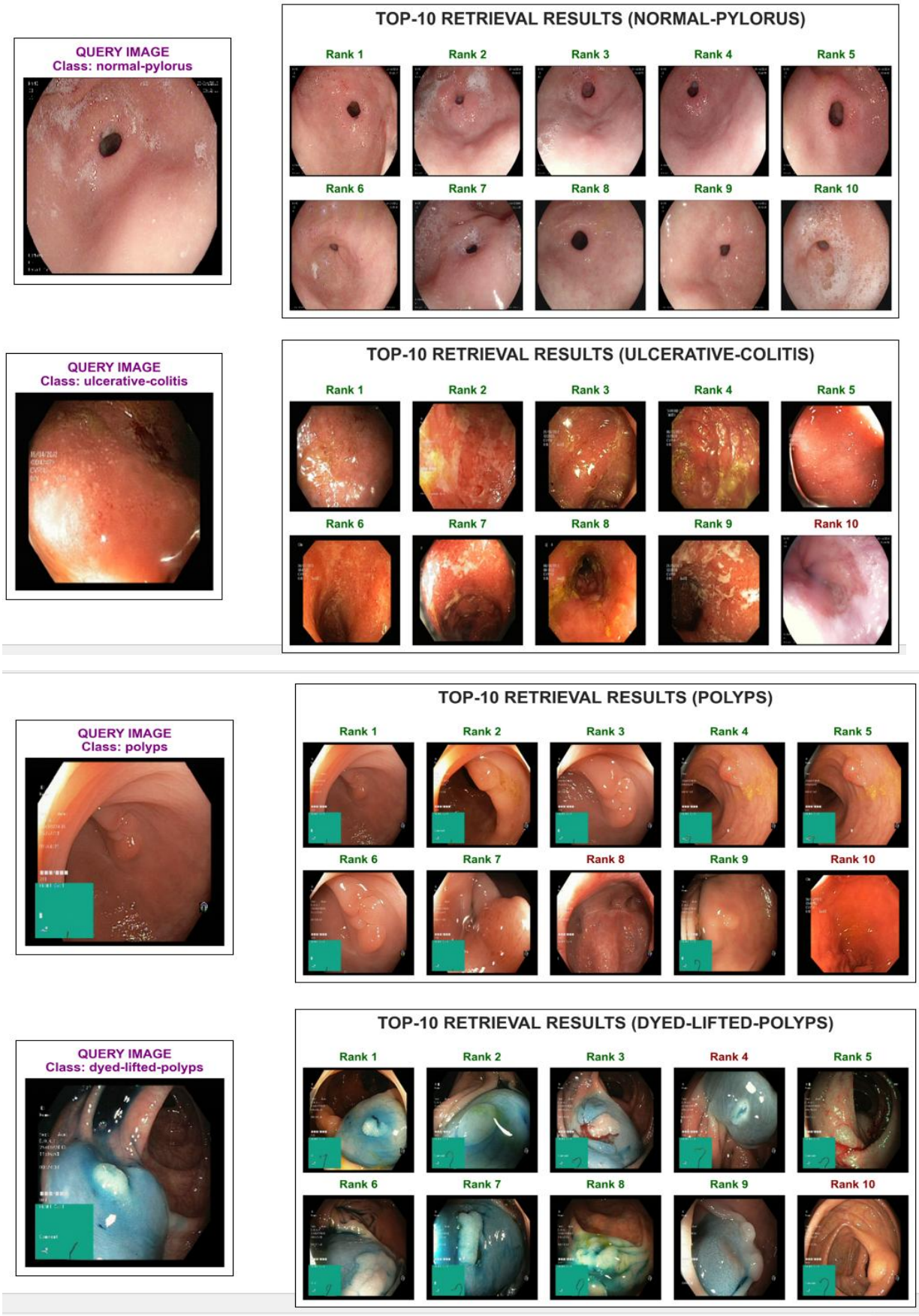


Fig. 10. Top-10 Medical Image Retrieval Results for Normal-Pylorus, Ulcerative-Colitis, Polyps, and Dyed-Lifted-Polyps Query Images

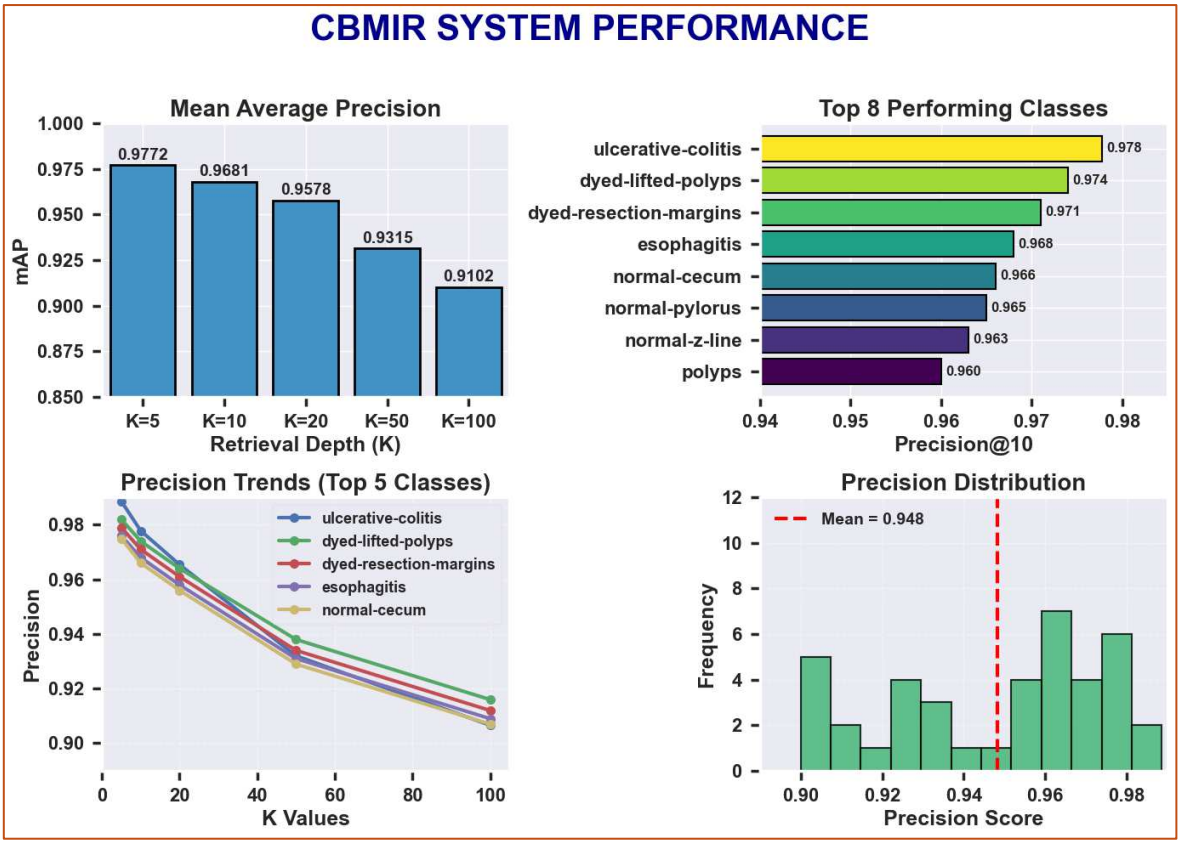


Fig. 11. Performance Analysis of the Proposed CBMIR Framework on the Kvasir Dataset

4.1.3 Quantitative Evaluation of Retrieval Performance

The performance of the proposed CBMIR model based on the Kvasir dataset is shown below in Figure 11. As shown through the MAP results in the figure, it can be said that the developed system exhibits a high degree of image retrieval performance as the system managed to score a MAP value of 0.9772 at K=5, which then decreased gradually until reaching a MAP value of 0.9102 at K=100. It clearly shows that the proposed method is capable of delivering highly relevant and accurate image retrieval irrespective of the value of K. Furthermore, the results of Precision@10 analysis on each class indicate that all the eight classes have excellent retrieval ability, where the classes such as ulcerative-colitis (0.978), dyed-lifted-polyps (0.974), and dyed-resection-margins (0.971) have obtained the best precision scores. From the precision curve of the top classes, it is clear that good retrieval is maintained at different levels of search depths. Moreover, the precision distribution histogram reveals that the majority of precision values are distributed in the range of 0.95 to 0.98 and yields an average precision value of about 0.948. These findings indicate that the proposed model consistently exhibits efficient performance in terms of medical image retrieval. The use of SPP-DenseNet121, CG-MViT, DMCCA-based feature combination, and PRF contributes significantly to this high retrieval accuracy by capturing distinctive features in medical images.

4.1.4 Class-wise Precision Analysis at Various Retrieval Depths

The class-wise precision performance of the CBMIR model is illustrated in Table III at various retrieval levels (P@5, P@10, P@20, P@50, and P@100), wherein P@K denotes the precision obtained through top-K retrieved images. The results indicate that the framework is able to deliver high levels of precision across all the eight classes of the Kvasir database. Among the examined categories, ulcerative-colitis has the best precision scores, with 98.86%, 97.78%, and 96.54% obtained at P@5, P@10, and P@20, respectively. In addition, dyed-lifted-polyps and dyed-resection-margins also provide very high retrieval accuracy with precision scores greater than 96 % for shallow retrieval depths. With an increase in retrieval depth from 5 to 100 images, there is a gradual decrease in precision for all classes due to the possible presence of fewer relevant images in large sets. Despite this, the precision values remain remarkably high and maintaining values which is higher than 90 percent for P@100.

Table III Class-wise Precision Performance of the Proposed CBMIR Framework at Different Retrieval Depths

Class	P@5 (%)	P@10 (%)	P@20 (%)	P@50 (%)	P@100 (%)
dyed-lifted-polyps	98.20	97.40	96.40	93.80	91.60
dyed-resection-margins	97.90	97.10	96.10	93.40	91.20
esophagitis	97.60	96.80	95.80	93.10	90.90
normal-cecum	97.50	96.60	95.60	92.90	90.70
normal-pylorus	97.40	96.50	95.50	92.80	90.60
normal-z-line	97.30	96.30	95.30	92.70	90.50
polyps	97.00	96.00	95.00	92.30	90.00
ulcerative-colitis	98.86	97.78	96.54	93.20	90.66
mAP (Mean)	**97.72**	**96.81**	**95.78**	**93.15**	**91.02**

Moreover, the mean average precision scores of 97.72%, 96.81%, 95.78%, 93.15%, and 91.02% corresponding to P@5, P@10, P@20, P@50, and P@100 reflect the consistency and efficacy of the developed search system. The findings validate the effectiveness of incorporating SPP-DenseNet121, CG-MViT, DMCCA, and PRF in accurately retrieving similar gastrointestinal image content based on visual appearance and semantics.

4.1.5 Ablation Analysis of the Proposed CBMIR Framework

The ablation study performed to assess the role of each individual major component of the suggested CBMIR architecture at various retrieval depths is shown in Table IV. Five different architectures are tested here, which include SPP-DenseNet121, CG-MViT, SPP-DenseNet121 + CG-MViT, SPP-DenseNet121 + CG-MViT + DMCCA, and the complete suggested architecture. It is evident from the table that adding components to the system leads to an improvement in retrieval performance. In individual comparison, CG-MViT performs better than SPP-DenseNet121 and shows the significance of feature learning using global context. When the SPP-DenseNet121 and CG-MViT models are combined, performance is improved due to the combination of local and global features of images. The use of DMCCA feature fusion contributes to better performance through the reduction of redundancy. Finally, the complete proposed framework has achieved the best retrieval performance for all retrieval depths, with mean precision scores of 97.72%, 96.81%, 95.78%, 93.15%, and 91.02% for P@5, P@10, P@20, P@50, and P@100, respectively. These observations indicate that each building block performs positively towards the medical image retrieval task, and the integration of multi-scale feature extraction, global contextual learning, DMCCA, and query refinement successfully generates higher accuracy when retrieving images from the Kvasir dataset.

The class-wise ablation study results of the proposed CBMIR model for various retrieval levels are provided in Table V. It can be clearly seen from the ablation study results that all the modules are contributing positively towards the retrieval mechanism, which is producing discriminative feature representations for better image retrieval performance.

Table IV Ablation Study Comparison of Different Framework Configurations at Various Retrieval Depths (mAP %)

Method	P@5 (%)	P@10 (%)	P@20 (%)	P@50 (%)	P@100 (%)
SPP-DenseNet121	92.18	91.01	89.99	88.25	85.12
CG-MViT	94.15	92.93	92.09	89.87	86.69
SPP-DenseNet121 + CG-MViT	95.40	94.63	93.59	91.13	87.92
SPP-DenseNet121 + CG-MViT + DMCCA	96.70	95.91	94.93	92.08	89.09
Proposed Framework	**97.72**	**96.81**	**95.78**	**93.15**	**91.02**

Table V Class-wise Ablation Study Comparison of Different Framework Configurations Across Retrieval Depths (%)

Class	Retrieval Depth	SPP-DenseNet121	CG-MViT	SPP-DenseNet121 + CG-MViT	SPP-DenseNet121 + CG-MViT + DMCCA	Proposed
Dyed-Lifted-Polyps	P@5	92.40	94.30	95.80	97.10	**98.20**
	P@10	91.60	93.50	95.20	96.50	**97.40**
	P@20	90.50	92.60	94.10	95.50	**96.40**
	P@50	88.95	90.64	91.92	92.86	**93.80**
	P@100	86.48	88.02	89.26	90.42	**91.60**
Dyed-Resection-Margins	P@5	92.10	94.10	95.60	96.90	**97.90**
	P@10	91.30	93.20	94.90	96.20	**97.10**
	P@20	90.20	92.30	93.80	95.20	**96.10**
	P@50	88.58	90.22	91.48	92.44	**93.40**
	P@100	86.06	87.62	88.88	90.04	**91.20**
Esophagitis	P@5	91.80	93.80	95.30	96.60	**97.60**
	P@10	91.00	92.90	94.60	95.90	**96.80**
	P@20	89.90	92.00	93.50	94.90	**95.80**
	P@50	88.32	89.96	91.18	92.16	**93.10**
	P@100	85.76	87.34	88.56	89.72	**90.90**
Normal-Cecum	P@5	91.70	93.70	95.20	96.50	**97.50**
	P@10	90.80	92.70	94.40	95.70	**96.60**
	P@20	89.70	91.80	93.30	94.70	**95.60**
	P@50	88.14	89.78	91.02	91.98	**92.90**
	P@100	85.58	87.14	88.36	89.52	**90.70**
Normal-Pylorus	P@5	91.60	93.60	95.10	96.40	**97.40**
	P@10	90.70	92.60	94.30	95.60	**96.50**
	P@20	89.60	91.70	93.20	94.60	**95.50**
	P@50	88.05	89.66	90.92	91.86	**92.80**
	P@100	85.48	87.04	88.26	89.42	**90.60**
Normal-Z-Line	P@5	91.50	93.50	95.00	96.30	**97.30**
	P@10	90.50	92.40	94.10	95.40	**96.30**
	P@20	89.40	91.50	93.00	94.40	**95.30**
	P@50	87.92	89.56	90.82	91.76	**92.70**
	P@100	85.34	86.92	88.14	89.30	**90.50**
Polyps	P@5	91.20	93.20	94.70	96.00	**97.00**
	P@10	90.20	92.10	93.80	95.10	**96.00**
	P@20	89.10	91.20	92.70	94.10	**95.00**
	P@50	87.54	89.16	90.46	91.34	**92.30**
	P@100	84.86	86.44	87.68	88.84	**90.00**
Ulcerative-Colitis	P@5	93.10	95.00	96.50	97.80	**98.86**
	P@10	92.00	94.00	95.70	96.90	**97.78**
	P@20	91.60	93.60	95.10	96.00	**96.54**
	P@50	88.46	89.98	91.26	92.22	**93.20**
	P@100	85.42	86.98	88.22	89.48	**90.66**

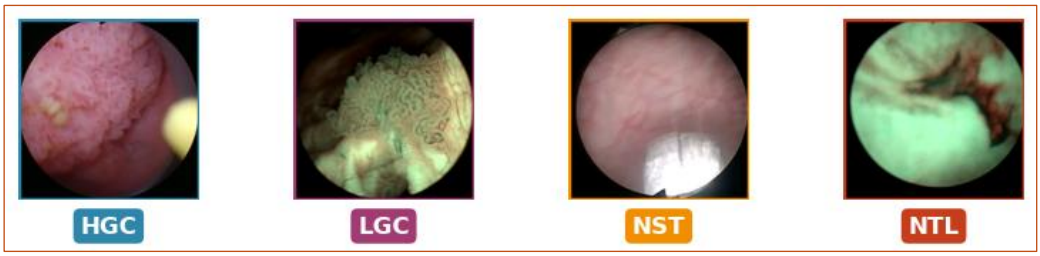


Fig.12. Sample images of Endoscopic Bladder Tissue Dataset

4.2 Dataset Description (Endoscopic Bladder Tissue Dataset)

The effectiveness of the proposed model is assessed via the Endoscopic Bladder Tissue Dataset, a publicly available dataset designed to assist in the identification and segmentation of bladder tissues in the Trans Urethral Resection of Bladder Tumors (TURBT). This dataset comprises 1,754 endoscopic images that were gathered from 23 patients, and these images were captured using two imaging techniques: White Light Imaging (WLI) and Narrow Band Imaging (NBI). All images were appropriately labeled according to their histopathology results.

The dataset can be classified into four clinically important tissue types, which include High-Grade Cancer (HGC), Low-Grade Cancer (LGC), No Tumor Lesion (NTL), and Non-Suspicious Tissue (NST). They represent different clinical levels of the cancer condition as well as different tissue conditions. As there are many different modalities and different visual representations of tissue, there is high visual diversity which makes it a tough challenge to test the suggested image retrieval framework. The sample image of this dataset is illustrated in Figure 12.

4.2.1 Effectiveness of Image Preprocessing and Contrast Enhancement

Figure 13 shows the impact of the proposed preprocessing technique on some endoscopic bladder tissues images and their respective pixel intensity histograms.

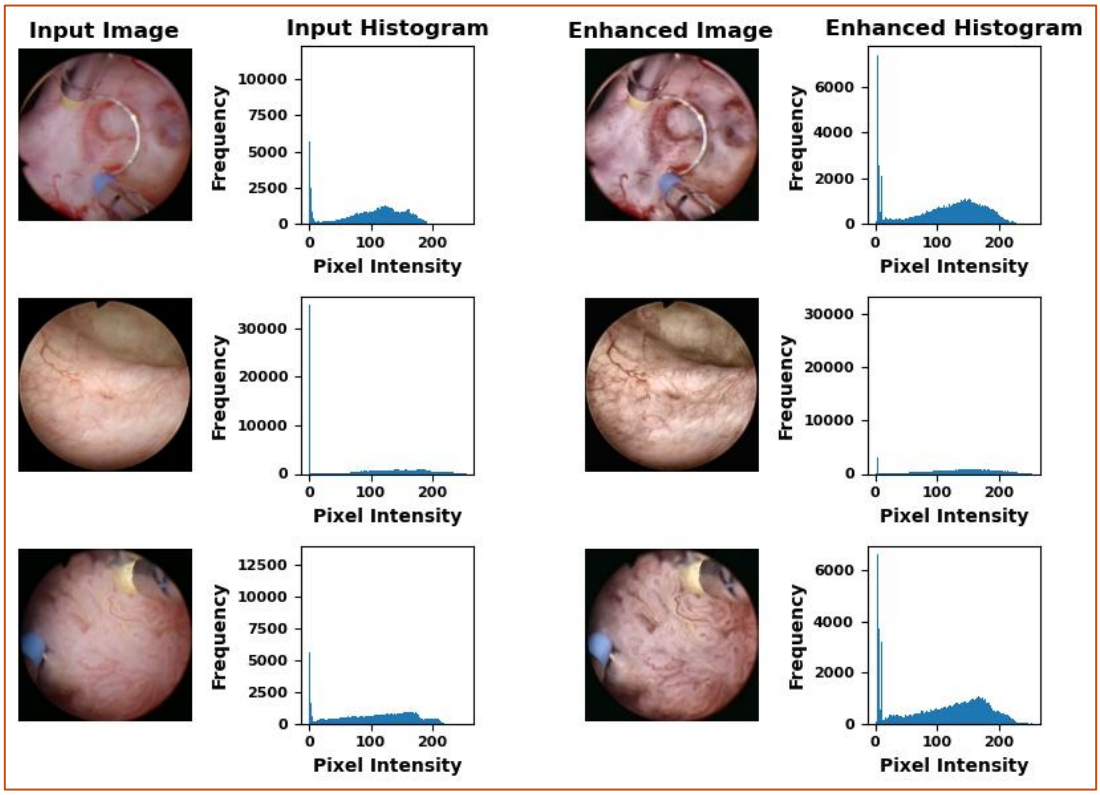


Fig. 13. Visual and Histogram-Based Analysis of Adaptive Contrast Enhancement on Endoscopic Bladder Tissue Images

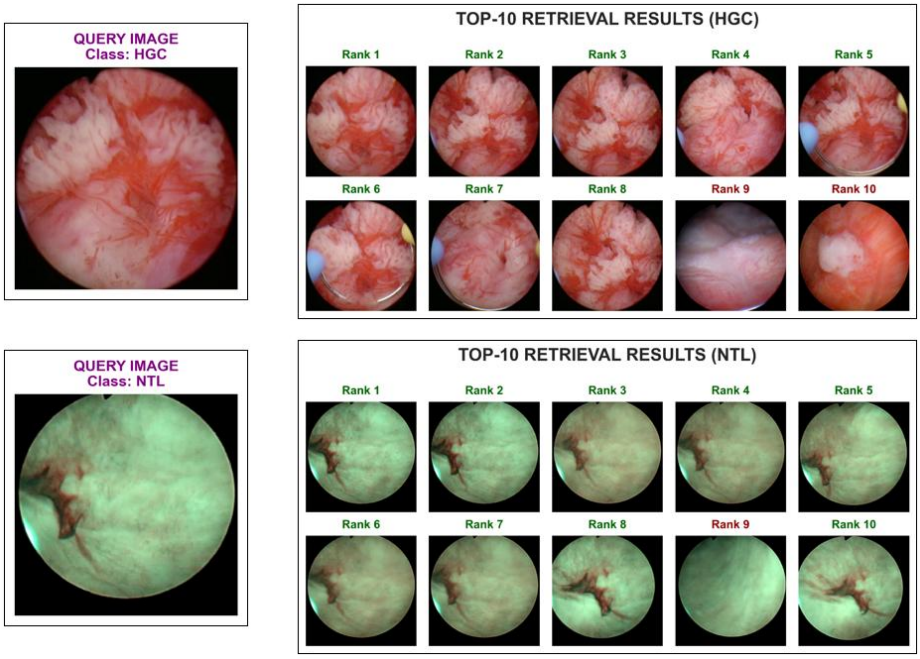


Fig. 14. Top-10 Retrieval Results for High-Grade Cancer (HGC) and No Tumor Lesion (NTL) Query Images

While the first two columns include the original input images and their histograms, the last two columns show the enhanced images using adaptive contrast enhancement and their respective enhanced histograms. From these figures, it can be seen that preprocessing process enhances the visibility of images by improving tissue contrast, illumination conditions, and detecting tissue features like blood vessels and lesions. The histogram distribution obtained after the image enhancement phase shows better distribution of intensity values within the dynamic range and implying better utilization of the contrast and brightness of the images. This leads to better distinction of clinically relevant attributes of the bladder tissues and consequently better feature extraction and learning for the subsequent retrieval process. This illustrates the effectiveness of the preprocessing approach and the resulting quality of the inputs for the CBMIR system.

4.2.2 Qualitative Analysis of Retrieval Results

The Figure 14 shows the qualitative results obtained in retrieving the most relevant images in the HGC and NTL category images through the proposed CBMIR approach. This is achieved by showing the top 10 closest related images retrieved from the dataset for the query images. It should be noted that the majority of the retrieved images have visual features that are highly similar to the query image.

Within the HGC class, all the retrieved images exhibit almost the same tumor structure, vessels, texture, and color distribution in relation to high grade cancer tissue. Similarly, in relation to the non-tumor tissue appearance, the retrieved images from the NTL query image are consistent with almost the same visual properties as those exhibited by the non-tumor tissue class. Although there is some slight deviation in some low-ranked images and they do not significantly affect the relevance of the overall image set retrieved from the respective query. This proves that the approach of combining SPP-DenseNet121, CG-MViT, DMCCA-based feature fusion, and query refinement is successful in capturing distinctive bladder tissue features in images.

4.2.3 Quantitative Analysis of Retrieval Performance

Figure 15 shows the overall retrieval effectiveness of the proposed CBMIR framework by applying different measures. From the mean average precision (mAP) performance results, it can be seen that the CBMIR framework attains an excellent retrieval accuracy of 0.926 at K = 5 and 0.907 at K = 10, implying that the framework retrieves highly relevant images from the top-ranked retrieval

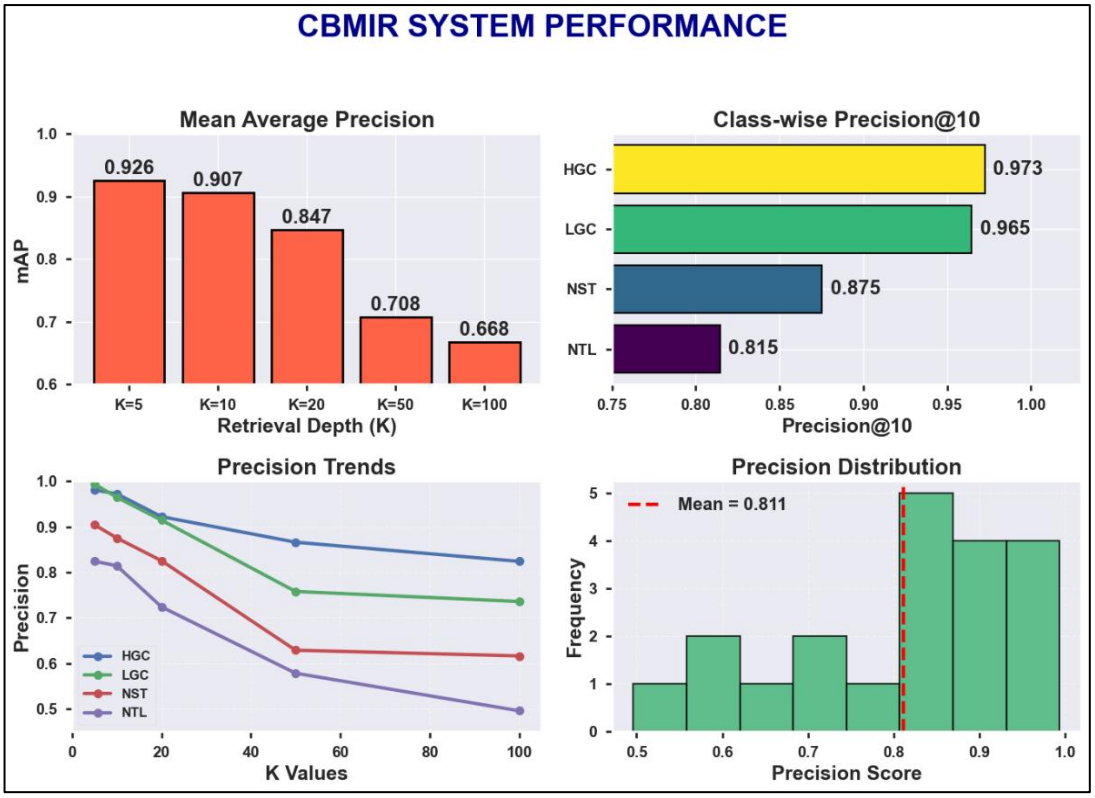


Fig. 15. Performance Evaluation of the Proposed CBMIR Framework on the Endoscopic Bladder Tissue Dataset

results. With the increase in retrieval depth to K = 20, K = 50, and K = 100, the mAP progressively reduces to 0.847, 0.708, and 0.668, respectively. HGC and LGC display superior performance in terms of Precision@10, where their precision is recorded as 0.973 and 0.965, respectively. These higher values suggest that HGC and LGC have a high level of discriminating power against non-cancerous tissue patterns. Conversely, NST and NTL display lower levels of Precision@10 and recording a value of 0.875 and 0.81 respectively. The trend graph for the precision further shows that both HGC and LGC have a stable pattern regarding retrieval effectiveness at various retrieval depths while NST and NTL show some degree of fall with the increasing numbers of images retrieved. On top of that, the histogram for precision distribution indicates an average precision rate of around 0.811, which again reflects the stability in the retrieval effectiveness of the endoscopic images. This clearly proves that the use of hybrid features with DMCCA fusion and query refinement has produced effective results for retrieval.

4.2.4 Class-wise Retrieval Precision Analysis

The class-wise precision rates of the proposed CBMIR method at various retrieval depths for Endoscopic Bladder Tissue Dataset is presented in Table VI. It can be seen from the table that the highest performance of the proposed technique is shown by LGC and HGC classes with precision rates of 99.40% and 98.09% at P@5, respectively.

Table VI Class-wise Precision Performance of the Proposed CBMIR Framework at Different Retrieval Depths on the Endoscopic Bladder Tissue Dataset

Class	P@5 (%)	P@10 (%)	P@20 (%)	P@50 (%)	P@100 (%)
HGC	98.09	97.25	92.23	86.62	82.41
LGC	99.40	96.47	91.47	75.80	73.60
NST	90.45	87.55	82.55	62.90	61.65
NTL	82.45	81.45	72.42	57.85	49.60
mAP (Mean)	**92.60**	**90.68**	**84.67**	**70.79**	**66.82**

Furthermore, the NST category also shows high accuracy in retrieving while the NTL category shows relatively low accuracy owing to the higher similarity in their visual characteristics. The level of precision gradually decreases when the depth of retrieval increases from 5 to 100 for all categories, which is natural as higher retrieval leads to higher chances of having irrelevant samples. The high mean average precision scores achieved by 92.60%, 90.68%, 84.67%, 70.79%, and 66.82% for P@5, P@10, P@20, P@50, and P@100 indicate that the suggested system preserves its robust image retrieval performance especially for high-ranked image results. These results clearly show that the suggested framework is capable of retrieving medically relevant images of bladder tissues from various tissue classes.

4.2.5 Ablation Analysis of the Proposed CBMIR Framework

The class wise ablation study is shown in Table VII and Table VIII shows the comparison between the mAP values at different retrieval depths of various configurations of the proposed CBMIR framework. It can be seen that the retrieval accuracy increases gradually with an increase in the number of the components included in the proposed CBMIR framework. Although the CG-MViT performs better than the SPP-DenseNet121 architecture individually, including both in the framework improves the effectiveness of image retrieval.

Including DMCCA for feature fusion brings about further improvement in retrieval accuracy by creating more discriminative hybrid features. The proposed framework produces the best retrieval accuracy at all retrieval depths, with mAP values of 92.60% and 66.82% at P@5 and P@100, respectively.

Table VIII Class-wise Ablation Study Comparison of Different Framework Configurations Across Multiple Retrieval Depths (%) on the Endoscopic Bladder Tissue Dataset

Class	Retrieval Depth	SPP-DenseNet121	CG-MViT	SPP-DenseNet121 + CG-MViT	SPP-DenseNet121 + CG-MViT + DMCCA	Proposed
HGC	P@5	91.65	93.84	95.72	97.18	**98.09**
	P@10	90.84	92.96	94.76	96.15	**97.25**
	P@20	85.36	87.42	89.28	90.74	**92.23**
	P@50	79.54	81.63	83.45	85.14	**86.62**
	P@100	75.26	77.54	79.46	80.93	**82.41**
LGC	P@5	92.84	95.23	96.87	98.12	**99.40**
	P@10	90.42	92.58	94.18	95.41	**96.47**
	P@20	84.72	86.95	88.54	89.96	**91.47**
	P@50	69.22	71.58	73.42	74.61	**75.80**
	P@100	67.21	69.45	71.22	72.46	**73.60**
NST	P@5	84.28	86.45	87.96	89.14	**90.45**
	P@10	81.24	83.65	85.21	86.46	**87.55**
	P@20	75.28	77.54	79.46	81.12	**82.55**
	P@50	56.45	58.84	60.32	61.54	**62.90**
	P@100	54.58	56.84	58.63	60.14	**61.65**
NTL	P@5	76.52	78.64	79.85	81.23	**82.45**
	P@10	74.13	76.52	78.25	79.84	**81.45**
	P@20	65.85	68.21	69.58	71.08	**72.42**
	P@50	51.71	53.86	55.42	56.63	**57.85**
	P@100	43.92	45.86	47.48	48.61	**49.60**

Table VII Mean Average Precision (mAP %) Comparison of Different Framework Configurations at Various Retrieval Depths

Retrieval Depth	SPP-DenseNet121	CG-MViT	SPP-DenseNet121 + CG-MViT	SPP-DenseNet121 + CG-MViT + DMCCA	Proposed
P@5	86.32	88.54	90.10	91.42	**92.60**
P@10	84.16	86.43	88.10	89.47	**90.68**
P@20	77.80	80.03	81.72	83.23	**84.67**
P@50	64.23	66.48	68.15	69.48	**70.79**
P@100	60.24	62.42	64.20	65.54	**66.82**

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This research proposed a new CBMIR framework by considering multiple scales feature extraction, context learning of images, discriminative feature fusion through DMCCA, and query expansion through the PRF technique. In addition, this proposed framework used complementary approaches of local and global feature learning to consider different features of an image, whereas DMCCA is used to produce discriminative hybrid features for retrieval purposes. Moreover, PRF based query expansion helps to narrow down the semantic gap between visual appearance and medical interpretation. Experimental outcomes on the Kvasir dataset yielded mAPs of 97.72%, 96.81%, 95.78%, 93.15%, and 91.02% for retrieval depths of P@5, P@10, P@20, P@50, and P@100, respectively. Likewise, in the case of the Endoscopic Bladder Tissue dataset, the developed approach produced mAPs of 92.60%, 90.68%, 84.67%, 70.79%, and 66.82% for the same retrieval depths. These positive results were again validated by the ablation studies carried out for the qualitative retrievals. Overall, it can be concluded from these results that the proposed approach is not only robust but also scalable and has tremendous potential for the purposes of computer-aided diagnosis, case-based reasoning, and clinical decision support.

Moreover, it might be helpful to implement new approaches for feature extraction from raw data by relying on self-supervised training or foundation models to minimize dependence on annotated medical datasets and make the proposed approach more generalizable. Finally, lightweight model architectures and efficient indexing approaches might be developed for real-time retrieval in large-scale hospital information systems. Additionally, adaptive techniques for query expansion using attention-based mechanisms and relevance feedback with clinician interaction may also be explored to improve the accuracy of the search process. Lastly, this suggested approach may be tested with a higher number of patients and expanded to other endoscopic/medical image datasets, like cystoscopy videos, histopathology, MRI, CT scans, and ultrasounds.

Acknowledgement:

The authors thank the International Research Centre of Kalasalingam Academy of Research and Education, Tamil Nadu, India for permitting the use of the computational facilities available in the Biomedical Research and Diagnostic Techniques Development Centre.

Funding:

There is no Research Funding for this article.

Conflicts of interest:

All authors declare that they have no conflicts of interest.

Availability of data and material:

The data set Kvasir is available in https://www.kaggle.com/datasets/alvikhan/kvasirv1

and Endoscopic Bladder Tissue Dataset is available in https://www.kaggle.com/code/xokent/endoscopic-bladder-tissue-classification.

Code availability:

PYTHON code is available.

Authors' contributions:

Conceptualization: Mathana Gopal Arulsamy; *Data curation:* Md. Jakir Hossen, Wai Kit Wong; *Formal Analysis:* Mathana Gopal Arulsamy; *Investigation:* Mathana Gopal Arulsamy; *Methodology:* Mathana Gopal Arulsamy; *Project administration:* Pallikonda Rajasekaran Murugan, Md. Jakir Hossen; *Supervision:* Pallikonda Rajasekaran Murugan; *Validation:* Poh Kiat Ng, M.Gomathy Nayagam; *Visualization:* Wai Kit Wong, Poh Kiat Ng, M.Gomathy Nayagam; *Writing – original draft:* Mathana Gopal Arulsamy; *Writing – review & editing:* Pallikonda Rajasekaran Murugan, Poh Kiat Ng, Wai Kit Wong, M.Gomathy Nayagam.

REFERENCES

1. Owais, Muhammad, Muhammad Arsalan, Jiho Choi, and Kang Ryoung Park. "Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence." *Journal of clinical medicine* 8, no. 4 (2019): 462.
2. Yildirim, Muhammed. "Content-based image retrieval and image classification system for early prediction of bladder cancer." *Diagnostics* 14, no. 23 (2024): 2637.
3. Agrawal, Shubham, Aastha Chowdhary, Saurabh Agarwala, Veena Mayya, and Sowmya Kamath S. "Content-based medical image retrieval system for lung diseases using deep CNNs." *International Journal of Information Technology* 14, no. 7 (2022): 3619-3627.
4. Ahmed, Ali. "Implementing relevance feedback for content-based medical image retrieval." *IEEE Access* 8 (2020): 79969-79976.
5. Srivastava, Divya, Shashank Sheshar Singh, Bakthula Rajitha, Madhushi Verma, Manjit Kaur, and Heung-No Lee. "Content-based image retrieval: A survey on local and global features selection, extraction, representation, and evaluation parameters." *IEEE Access* 11 (2023): 95410-95431.
6. Samantaray, Aswini Kumar, and Amol D. Rahulkar. "Content Based Medical Image Retrieval." In *Feature Extraction in Medical Image Retrieval: A New Design of Wavelet Filter Banks*, pp. 1-24. Cham: Springer Nature Switzerland, 2024.
7. Tyagi, Surbhi, Praveen Shukla, Partap Singh, and Hitesh Tekchandani. "Enhanced content-based image retrieval via hybrid color, texture, and deep learning features." *Scientific Reports* (2026).
8. Khokher, Amandeep, and Rajneesh Talwar. "A fast and effective image retrieval scheme using color-, texture-, and shape-based histograms." *Multimedia Tools and Applications* 76, no. 20 (2017): 21787-21809.
9. Samantaray, Aswini Kumar, and Amol D. Rahulkar. "Content Based Medical Image Retrieval." In *Feature Extraction in Medical Image Retrieval: A New Design of Wavelet Filter Banks*, pp. 1-24. Cham: Springer Nature Switzerland, 2024.
10. Garg, Meenakshi, and Gaurav Dhiman. "A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants." *Neural Computing and Applications* 33, no. 4 (2021): 1311-1328.
11. Shetty, Rani, Vandana S. Bhat, and Jagadeesh Pujari. "Content-based medical image retrieval using deep learning-based features and hybrid meta-heuristic optimization." *Biomedical Signal Processing and Control* 92 (2024): 106069
12. Chughtai, Iqra Toheed, Asma Naseer, Maria Tamoor, Saara Asif, Mamoona Jabbar, and Rabia Shahid. "Content-based image retrieval via transfer learning." *Journal of Intelligent & Fuzzy Systems* 44, no. 5 (2023): 8193-8218.
13. Palla, Monica, and Renu Karra. "A Comparative Study of Pre-Trained CNN Models with Transfer Learning for Content-Based Image Retrieval." *Engineering, Technology & Applied Science Research* 15, no. 4 (2025): 25820-25826.
14. Ahmed, Ali, Alaa Omran Almagrabi, and Ahmed Hamza Osman. "Pre-trained convolution neural networks models for content-based medical image retrieval." *Int. J. Adv. Appl. Sci.* 9 (2022): 12.

15. Rani, Sunita, Geeta Kasana, and Shalini Batra. "An efficient content based image retrieval framework using separable CNNs." *Cluster Computing* 28, no. 1 (2025): 56.
16. Susmitha, Arvapalli Sai, and Vinay P. Namboodiri. "Analysis of transformers for medical image retrieval." In *Medical Imaging with Deep Learning*. 2024.
17. Kumar, Mehul, Rhythumwinder Singh, and Prerana Mukherjee. "VTHSC-MIR: Vision Transformer Hashing with Supervised Contrastive learning based medical image retrieval." *Pattern Recognition Letters* 184 (2024): 28-36.
18. Vijila, J., and Godfrey Winster Sathianesan. "Interactive Content-based Image Retrieval (Interactive CBIR): Hybrid Transformer-Compressed Convolutional Model for Deep Feature Extraction." In *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 103-109. IEEE, 2025.
19. Abderrahim, Marwa, Abir Baâzaoui, and Walid Barhoumi. "Survey on Multi-Source Medical Imaging Fusion for Classification and Retrieval: Current Status and Available Datasets." *SN Computer Science* 6, no. 7 (2025): 773.
20. Zhang, Rongchen. "An advancing framework for complex fabric image retrieval: from multi-scale feature fusion to efficient attention-based matching." (2025).
21. Ahmed, Ali, Alaa Omran Almagrabi, and Omar M. Barukab. "A Content-Based Medical Image Retrieval Method Using Relative Difference-Based Similarity Measure." *Intelligent Automation & Soft Computing* 37, no. 2 (2023).
22. Djoudi, Kaouthar, Zaia Alimazighi, and Badiâa Dellal Hedjazi. "Information retrieval with query expansion and re-ranking: a survey." In *IET Conference Proceedings CP906*, vol. 2024, no. 34, pp. 114-119. Stevenage, UK: The Institution of Engineering and Technology, 2024.
23. Ahmed, Ikhlaq, Naima Iltaf, Rabia Latif, Nor Shahida Mohd Jamail, and Zafran Khan. "Dual Modality Reverse Reranking (DM-RR) Based Image Retrieval Framework." *IEEE Open Journal of the Industrial Electronics Society* 5 (2024): 886-897.
24. Badanidiyoor, Ashwath Rao, Vidya Harekala Chandrashekara, and Naravi Gopalakrishna Kini. "Evaluation of Local Texture Features in Content-based Medical Image Retrieval Task." *Engineering Letters* 34, no. 1 (2026).
25. Singh, Mona, and Manoj Kumar Singh. "Content-Based Gastric Image Retrieval Using Fusion of Deep Learning Features with Dimensionality Reduction." *SN Computer Science* 6, no. 2 (2025): 184.
26. Humadi, Aqeel M., Mehdi Sadeghzadeh, Hameed A. Younis, and Mahdi Mosleh. "Adaptive Two-Tier Deep Learning for Content-Based Image Retrieval and Classification With Dynamic Similarity Fusion." *IET Image Processing* 19, no. 1 (2025): e70192.
27. Kim, Kangsan, Junseok Park, Sang Hyun Kim, and Youngbae Hwang. "Improving image-retrieval performance of foundation models in gastrointestinal endoscopic images." *Frontiers in Medicine* 12 (2025): 1727884.
28. Ahmed, Ali, and Sharaf J. Malebary. "Query expansion based on top-ranked images for content-based medical image retrieval." *IEEE Access* 8 (2020): 194541-194550.
29. Rani, Sunita, Shalini Batra, and Geeta Kasana. "A content-based image retrieval approach utilizing optimized intermediate layer features." *Cluster Computing* 29, no. 5 (2026): 275.
30. Ahmed, Ali, Alaa Omran Almagrabi, and Ahmed Hamza Osman. "Pre-trained convolution neural networks models for content-based medical image retrieval." *Int. J. Adv. Appl. Sci.* 9 (2022): 12.
31. Alzahrani, Ahmad A., Ali Ahmed, and Alisha Raza. "Content-based medical image retrieval method using multiple pre-trained convolutional neural networks feature extraction models." *Int. J. Adv. Appl. Sci* 11, no. 6 (2024): 170-177.
32. Ahmed, Ali, Alaa Omran Almagrabi, and Omar M. Barukab. "A Content-Based Medical Image Retrieval Method Using Relative Difference-Based Similarity Measure." *Intelligent Automation & Soft Computing* 37, no. 2 (2023).
33. M. G. Arulsamy, P. R. Murugan, A. P. Thiyagarajan and K. Ramaraj, "Feature Fusion Tactics in Gastrointestinal Endoscopic Image Retrieval Using Deep Features," *2023 International Conference on Energy, Materials and Communication Engineering (ICEMCE)*, Madurai, India, (2023): 1-6